Polymorphism and replication of heterochromatic repeats in the DNA of *Arabidopsis*

Jerry Davison

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2006

Program Authorized to Offer Degree: Biology

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Jerry Davison

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

Luca Comai

Reading Committee:

Arnold J. Bendich

Harvey D. Bradshaw

Luca Comai

Date: October 24, 2006

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature _____

Date 24 X 2006

University of Washington

Abstract

Polymorphism and replication of heterochromatic repeats in the DNA of *Arabidopsis*

Jerry Davison

Chair of the Supervisory Committee
Professor Luca Comai
Department of Biology

The composition of the individual eukaryote's genome and its variation within a species remain poorly defined. Even for a sequenced genome such as that of the model plant *Arabidopsis thaliana* accession Col-0, the large arrays of heterochromatic repeats are incompletely sequenced. We defined by flow cytometry a set of *A. thaliana* accessions that differ in measured nuclear genome size. Using these geographically separate populations, we assayed variation in the heterochromatic repeat arrays using two independent methods and identified substantial polymorphism among them, with variation by as much as a factor of two in the centromeric 180 bp repeat, in the 45S rDNA arrays and in the Athila retroelements. In the accession with highest measured genome size, Loh-0, we measured more than a two-fold increase in 5S RNA gene copies relative to Col-0; results from fluorescence in situ hybridization with 5S probes were consistent with the existence of size polymorphism between Loh-0 and Col-0 at the 5S loci. Comparative genomic hybridization results of Loh-0 and Col-0 did not support contiguous variation in copy number of protein-coding genes on the scale needed to explain their observed genome size difference. We developed a computational data model to test whether the variation we measured in the repeat fractions could account for the different genome sizes determined with flow cytometry, and found that this proposed relationship could account for about

50% of the variance in genome size among the accessions. Our finding of a negative relationship between measured genome size and the copy number of centromeric repeats was unexpected, as centromeres may make up the single largest repeat class in the genome. Heterochromatic repeats have been demonstrated to be much less than fully replicated in the endoreplicated cells of some taxa. We investigated whether under-replication of repeated heterochromatic sequences in the higher ploidy nuclei of Arabidopsis distorts our genome size measurements or our assessment of the amounts of heterochromatin in the accessions. We found the under-replication amounts we observed with flow cytometry to be insufficient by a factor of five to account for the negative relationship between measured genome size and centromeric repeat amount.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acknowledgements

I'd like to thank Luca Comai for welcoming me into his laboratory and for his advise and counsel during my graduate studies. I also thank Arnie Bendich and Toby Bradshaw for their service on my committee. Many, many thanks to all the members of the Comai lab. I sincerely and deeply thank the faculty and staff of the Botany and Biology departments for their great kindness, generosity and assistance during my tenure here. I gratefully acknowledge funding of scholarships from the Department of Botany, and the Cellular and Molecular Biology Training Grant (PHS NRSA T32 GM07270).

Chapter I. Introduction

Genome size differences are present at all taxonomic levels in eukaryotes (Comeron, 2001; Gregory and Hebert, 1999). In land plants, nuclear DNA content ranges from 100 million to 100 billion base pairs, three orders of magnitude from Arabidopsis to irises. Within a family 30-fold differences are common, and within a genus, a 5-fold range is typical (Bennett, 1998). Genome size among South American populations of the wild peanut (*Arachis duranensis*) varies by 8% (Temsch and Greilhuber, 2001), and strains of laboratory mice vary by the same amount (Capparelli, 1997).

Significantly, differences in genome size among taxa have no correlation with organism complexity, a circumstance known historically as the C value problem (Petrov, 2001). For example, the 1C (haploid, or single complement) genome of *Homo sapiens* is 3.4 Gbp (billion base pairs of DNA), *Zea mays* 5.0 Gbp, and the protozoan *Amoeba dubia*, 670 Gbp, 200 times that of humans. The problem has been partly resolved with determination that the bulk of a genome's DNA does not code for protein. Heterochromatic DNA accounts for much or most of the differences in genome size among eukaryotes (Kidwell, 2002). A highly condensed chromatin, heterochromatin (Redi *et al.*, 2001), alternates in chromosomes with the open arrangement of euchromatin, where most of the active genes are located (Klug and Cummings, 2004).

Heterochromatin typically consists of sequences up to thousands of nucleotides long, often tandemly repeated (Bennetzen, 2000; CSHL, 2000). Even when non-coding, repeats in heterochromatin make up functional neighborhoods of chromosomes (Dillon and Festenstein, 2002); centromeres (Haupt *et al.*, 2001) and telomeres (Lohe and Hilliker, 1995) are major examples.

Intriguingly much of constitutive heterochromatin is made up of fragments or complete sequences of transposable elements (TEs) (Dimitri and Junakovic, 1999) which though silenced, can be activated (Feschotte *et al.*, 2002). Replication of TEs and their silencing is a mechanism known to change DNA

content — Vieira *et al.* (2002) showed in both *Drosophila melanogaster* and *D. simulans* that genome size varies among populations and is correlated with the TE copy number.

Significance of genome size variation

From the later 19th through the 20th centuries, questions about genome organization and its regulatory and evolutionary role have been taken up repeatedly. While the physical basis of inheritance and gene regulation have been clarified in time, complexities continue to surface and test understanding.

With the discovery of non-coding DNA, the lack of correlation between genome size and organismal sophistication is explained, but an increasingly elaborate genomic citizenry is being recognized (Kubis *et al.*, 1998; Puertas, 2002). Potentially expressed and troublesome fragments such as transposons exist, along with untranscribed control sequences like boundary elements (Ishii *et al.*, 2002), and transcribed but untranslated RNAi genes.

The structural organization of nuclear DNA is known to be a fundamental regulator of gene expression. In addition to discrete elements, the continuum of secondary and higher structure in chromatin fiber is now recognized as critically important in genome operation (Manuelidis, 1990; Jackson, 1991; Lamond and Earnshaw, 1998; Cremer and Cremer, 2001; Iborra and Cook, 2002). Changes in heterochromatin can modify gene expression. Position effect variegation occurs when a normally euchromatic region is juxtaposed with heterochromatin in a translocation event. Apparent stochastic spread of heterochromatin into a gene results in its variegated expression (Clark *et al.*, 1996).

Whether allelic incompatibilities *sensu* Dobzhansky-Muller are more significant in reproductive isolation, or chromosomal rearrangement and non-genic factors as argued by Bateson, Goldschmidt and others, is not agreed upon. Knowledge gained in the 100 years since the discussion started reveals more complexity than earlier supposed.

It is clear that gross chromosomal change can isolate conspecifics. Sharma *et al.* (2003) found reduced fitness among hybrids between chromosomal races of the Indian pygmy field mouse. *Mus terricolor I*, *II* and *III* are three chromosomal species with heterochromatin variations established in homozygous condition. Several researchers have shown that, in meiosis in F1 hybrids of closely related plant species, the number of bivalents formed at pachytene falls with increasing genome size difference between the parents (Levin 2002, pp. 74-76).

The relationship between heterochromatin polymorphism and disease in humans is a recurring topic. Atkin and Brito-Babapulle (1981) cited support for a link in two ways: individuals with heterochromatin polymorphism may be at risk for increased congenital abnormalities in their offspring, and polymorphism may be associated with chromosomal instability in somatic cells.

What mechanisms could be responsible for reduction in fitness in lineages from parents with different heterochromatin loci? One possibility is activation of transposable elements in the developing embryo in reaction to non-homology between homologous chromosomes. A poorly characterized but well-established genomic shock or stress occurs in the offspring of plants differing in ploidy levels, and in hybrids between congenerics. Barbara McClintock first used this term to explain certain atypical chromosomal events in plant reproduction (McClintock, 1984).

In meiosis as a rule, synapsis occurs along the length of every homologous chromosome. Dernburg *et al.* (1996) showed that in Drosophila the heterochromatin of homologs is paired from pachytene of prophase I until chromosome segregation. McClintock's detailed observations traced rearrangement and previously unrecorded interactions in the genetic material; among these were the non-homologous pairing (McClintock, 1930) and complete synapsis between homologous chromosomes carrying inversions and translocations (McClintock, 1933).

Can reproductive isolation result from accumulation of seemingly benign, non-coding changes, or do mechanisms exist to accommodate or counter them? Do changes in genome organization follow or precede speciation? There are no conclusive answers.

Variation in the repeated fraction

The fundamental mechanisms that generate and shape genomic diversity — mutation, recombination, selection and drift — were well known before the genomic era. Despite advances, the composition of a eukaryote species' genome along with its variation from individual to individual are still not well defined. The decreasing cost of newer DNA sequencing techniques makes possible the large-scale resequencing of DNA from multiple individuals, providing important information on single nucleotide polymorphisms and on small insertion-deletions (indels); but resequencing is ill-suited to describe a significant source of intraspecific diversity, variation in the copy number of genomic elements. Copy Number Variation, CNV, is defined (Freeman *et al.*, 2006) as deletions or duplications of any genomic elements, except transposons, greater than one thousand base pairs (bp). Perfect duplications of single copy sequences can escape detection, and indels are detected only by comparison with other genomes. Emerging research suggests that genic CNV contributes to major remodeling between species, and disease in humans (Aitman *et al.*, 2006; Fortna *et al.*, 2004; Sebat *et al.*, 2004; Freeman *et al.*, 2006).

Early protein electrophoresis studies demonstrated unexpected genic polymorphism within species, and chromosome structure research found large differences in total genome content across taxa. Reassociation kinetics assays have identified sizeable variation in the fast reannealing fraction of the nuclear genome, which corresponds to repeated sequences. In eukaryotes this fraction is substantial; highly and moderately repetitive sequences are each typically on the order of the size of the non-repetitive fraction (Lewin, 1997, pp. 654-655). Centromeres are familiar repeated structural elements, and the ribosomal RNA

genes are duplicated many times; both are in arrays maintained as heterochromatin. No accepted term is in use to define copy number variation in transposons, transposon-related, centromeric and ribosomal repeats. To facilitate this discussion, we will designate this latter type of variation as Repeat Number Variation (RNV). RNV can arise rapidly (Bennetzen, 2002), and multiple mechanisms have been proposed to account for it: non-homologous recombination, excision, duplication, inversion, transposition, and proliferative replication of retroelements (Jackson *et al.*, 1999). Stress of different types has been suggested to trigger remodeling that leads to substantial changes. The significance of CNV in non-genic elements is unclear — in the human population heterochromatin CNV has been reported both as general with no effect, and associated with disease (Park *et al.*,1998; Ji *et al.*, 2000; Yakin *et al.*, 2005).

RNV is hard to characterize. The larger repeat-rich sequences of the genome cannot be tiled into contigs without ambiguity, due to their repetitive nature, and gaps of uncertain but megabase size persist in the sequenced genomes' repeats, including the human, in particular in centromeres (Eichler *et al.*, 2004; TIGR, 2000). For that reason major repeats have been excluded from the definition of a sequenced genome (Wortman *et al.*, 2003). The uncertainty in the repeated component is illustrated by the status of the nuclear genome of the model organism Arabidopsis, one of the smallest in the vascular plants. The initial *Arabidopsis thaliana* genome sequence was announced by the Arabidopsis Genome Initiative (AGI) in 2000, with the 1C genome estimated to be 125 million base pairs (Mbp); 115 Mbp had been sequenced, with work continuing on the centromeres and 5S rDNA. Subtelomeric rDNA arrays on chromosomes 2 and 4 were not sequenced.

The sizes of all 5 centromeres were subsequently reassessed (Kumekawa *et al.*, 2000; Kumekawa *et al.*, 2001; Hosouchi *et al.*, 2002) with sequence analysis of BAC and YAC clones anchored in the pericentromeric regions followed by physical mapping of the centromeric gaps using restriction digests. The authors estimated the total extent of the genetically defined centromeres to

be 27 Mbp, three times the initial AGI estimate of 8.7 Mbp, placing the total genome size near 146 Mbp. Similar conclusions were reached by Bennett *et al.* (2003) who using flow cytometry of fluorophore-dyed nuclei found a content larger than estimated by the AGI, and proposed that the sizes of the major repeats were considerably larger. Weighing reported sizes of the NOR rDNA and centromeres, the authors calculated the two were underestimated by 25 Mbp.

Even with this imprecision in the repeated fraction the *Arabidopsis thaliana* nuclear genome is one of the best characterized eukaryotic genomes, and provides an opportunity to better understand heterochromatin copy number variation in a model organism. Comparison of accessions with substantially different genome size can potentially reveal the major sequences subject to variation; the method we used to measure genome sizes is flow cytometry, based on the fluorescence of DNA-bound dye. It can be accurate and agrees well with an earlier standard method, Feulgen staining, but its limitations are not adequately appreciated (Suda, 2004). Chromatin condensation status substantially affects its measurements, and while chromatin's variability in response to conditions (pH and salinity for example) is well-documented (Giangarè *et al.*, 1989; Noirot *et al.*, 2000), developmental and evolutionary changes in chromatin are not well understood.

Uncertainty in copy number of the repeated fraction

We used published information to build Table 1.1; the sequenced and estimated sizes of several genomic classes in *A. thaliana* are presented. The sequenced accession Columbia (Col-0) has some 26,000 predicted protein-coding genes, each averaging 2,000 bp, with 65% in multiple copies. Over 5,600 transposable elements (TEs) concentrated in and near centromeres have been identified (AGI, 2000); Athila is the most abundant pericentromeric TE, making up more than a third of flanking sequences. The Arabidopsis genome is made up by five chromosomes.

The centromeres, only partly sequenced, are together estimated to be 8 to 28 Mbp. The ten telomeres make up 100 thousand base pairs (Kbp). Two regions of 45S ribosomal RNA-encoding DNA (the NOR loci) also not sequenced are distal on chromosomes 2 and 4 North. The 5S ribosomal RNA gene arrays have been identified near centromeres 4 and 5, and 3 in some accessions (Fransz *et al.*, 1998). Intergenic sequences comparable in total size to the genic genome fill out the approximately 150 Mbp.

We noted that the understanding of the repeated fraction of the genome is incomplete. To clarify that we briefly review the status of one of the smaller repeats, the 5S ribosomal RNA gene array. The individual 5S repeat unit is 497 bp (Cloix *et al.*, 2002), with a 121 bp RNA transcript, a component of the large ribosomal subunit (Szymanski *et al.*, 2003). There is variation in repeat lengths in the size of the repeat unit.

In a report often cited, Campell *et al.* (1992) estimated there to be around 1000 copies of the 5S rDNA unit in Arabidopsis accessions En-2, Col-0, An-1 and Nd-0. The authors estimated copy number by comparing the hybridization of labeled 5S units excised from a plasmid to known amounts of dot-blotted plasmid and genomic DNA. To convert copies per DNA amount to copies per genome they assumed a haploid genome size of 70 Mbp. Doubling this estimate of genome size to today's value near 150 Mbp gives approximately 2000 copies per haploid genome, or about 1 Mbp.

Fransz *et al.* (1998), using Fluorescence In Situ Hybridization (FISH) with 5S probes found 5S rDNA on chromosomes 3, 4, & 5 in Col-0, with polymorphism present among six accessions in three patterns. In all accessions assayed there are major loci in the short arm of chromosome 4 and the upper arm of chromosome 5. A minor locus is present in the lower arm of chromosome 5 in accessions WS, Col-0, L*er*, and C24, "though not always observed." A third major locus was found on chromosome 3 in L*er*, Col-0, Kas1 and Cvi; the position of the third locus is variable. The authors pointed out that the polymorphism implies that Campell's findings are not valid for all accessions.

Cloix *et al.* (2000) looked closely at 5S rDNA repeats in the left, or upper, arm of chromosome 5; they constructed a contig going through the 5S locus identified by Fransz *et al.* using YAC clones from the CIC library (Creusot *et al.*, 1995). They found three 5S blocks approximately 100, 30, and 150 Kbp long as determined by pulsed field electrophoresis, totaling 280 Kbp. With these data and Fransz *et al.*'s finding of three major loci and one minor 5S locus in the Col-0 genome and assuming equal size for each major locus, one can roughly estimate the full 5S complement as 280 Kbp/locus times 3.5 loci, about 1 Mbp in this accession.

The methods used in these papers provide the best available data on the locations and sizes of the 5S repeats; similar efforts producing physical and genetic maps across the genome guided the sequencing project (AGI 2000). The tools and strategies used in genome sequencing are well-documented (Fleischmann *et al.*, 1995; CESC, 1998; Myers *et al.*, 2000); the repeated sequences, whether tandem, interspersed or segmental, generate the greatest challenges.

The sequenced Arabidopsis 5S rDNA regions contain clone gaps, intervals in the genome that the available clones do not span (TAIR, 2006). They result from the difficulty of cloning highly complex and repeated sequences, which may be unstable in vectors. Discontinuities within clones, known as sequencing or finishing gaps, are present in clones that include the 5S RNA gene arrays; these gaps result from ambiguity in constructing a contig with repeated nucleotide sequences. Reads with identical sequences may be from clones that overlap in the genome or from completely separate ones. Methods have been developed to address this, for example by identifying single base pair differences among collections of repeats to aid in resolution of ambiguities (Arner *et al.*, 2006), but the additional costs, perceived lack of importance and often absence of practical method work against the full characterization of repeated loci.

The discrepancy between the 5S loci identified by FISH and other methods and their abundance in the sequenced genome can be noted with the

TAIR Sequence Viewer (TAIR, 2005). Using this software we probed the sequenced Arabidopsis genome with the shortest fragment of the 5S RNA gene the Sequence Viewer accepts (15 bp). There were 205 sequences found in chromosome 3, 10 in chromosome 4, and 74 in chromosome 5, giving an Arabidopsis 5S total near 150 Kbp. The TIGR BAC tiling path (TIGR, 2000) sets aside additional space for the 5S cluster on chromosome 4 (estimated at 650 Kbp) and the two 5S clusters on chromosome 5 (estimated at 260 Kbp). The gaps indicate that the ideal of the genomic era is not yet realized.

Table 1.1. Three estimates for the size of the *Arabidopsis thaliana* genome. Units are millions of base pairs (Mbp). Data sources are S: Sequencing, L: Literature, P: Physical mapping.

Sequence class	AGI (2000)	Source	Hosouchi et al. (2002)	Source	Bennett et al. (2003)	Source
Genes	51	S	—		—	
Intergenic DNA	59	S	—		—	
Centromeres	8	L	27	P	28	L
Distal rDNA	7	L	—		10 to 12	L
Estimated total	**125**		**146**		**147**	

Chapter II. Polymorphism of heterochromatic repeats

Introduction

We repeated a recent study surveying nuclear genome content in *A. thaliana* to begin our assessment of intraspecific variation in its heterochromatin. We chose for more detailed analysis those accessions (members of populations descending from different wild-collected plants, known also as ecotypes) that differed up to 20 Mbp. We characterized relative and absolute amounts of the heterochromatic repeat elements, and found large-scale polymorphism in the centromeric, pericentromeric and ribosomal repeats in Arabidopsis; we discounted the possibility of large euchromatic segmental duplications by comparative genomic hybridization. Finally we estimated the repeats' numerical contributions to total nuclear genome content.

Results

Genome size measurements

To assay variation in genome size and identify accessions with nuclear genome size larger and smaller than the sequenced accession Columbia (Col-0), we measured the genome size of 22 accessions acquired from the ABRC in Columbus, Ohio and the laboratory of Magnus Nordborg at the University of Southern California, using flow cytometry as described in Materials and methods. Our values are given in Table 2.1 and graphed in Figure 2.1. Nuclear DNA content in the 22 accessions ranged from 150 to 170 Mbp.

We measured the genome size of the sequenced accession Col-0 to be 157 Mbp (0.160 picogram), using commercially available alcohol-fixed chicken erythrocyte nuclei from Becton-Dickinson as the internal size standard, and taking the *Gallus gallus* 1C genome size to be 1150 Mbp. The draft sequence of the domestic chicken 1C GS is 1050 Mbp (Hillier *et al.*, 2004); previous estimates

range from 1150 to 1250 Mbp. Using this standard and DNA amount (1150 Mbp, 1.165 picogram), Bennett *et al.* (2003) reported the Col-0 accession GS to be 163.7 Mbp, 4% larger than our mean Col-0 value. Both values are more than 25% larger than the 125 Mbp estimated by the AGI (2000).

Schmuths *et al.* (2004) measured the genome size of 19 diploid Arabidopsis accessions using flow cytometry and *Raphanus sativus* (the cultivated radish) as an internal size standard. While this plant's genome has not been sequenced, interlaboratory estimates of its genome size have been made (Dolezel *et al.*, 1998); Schmuths *et al.* used the value from laboratory one, an *R. sativus* 1C estimate of 680 Mbp. The authors' measured value of the 1C Col-0 genome size is 202 Mbp, with the 19 accessions ranging from 202 to 221 Mbp.

Schmuths *et al.*'s largest and smallest measured values differ by 9% of the mean of all measurements while ours span 12% of the mean of our measurements. Three accessions are shared between the two sets; Col-0 has the smallest genome size in their measurements and is the fifth smallest in ours. The other shared accessions, Ag-0 and Tsu-0, are within 1% of the mean measured genome size in each set.

We selected five accessions identified in our survey of 22 for closer review and used these to investigate the sources of differential genome size in Arabidopsis. The results of the additional genome size assays are presented in Figure 2.2; the measured genome size (MGS) of each accession is presented relative to Col-0. Two accessions, Ta-0 and Br-0, have mean MGS smaller than the sequenced accession Col-0, and three, Is-0, TAMM-2 and Loh-0, are larger.

Measurements of the major repeats

We first considered the hypothesis that genome size differences could arise from RNV involving one or more of the major repeat families. To test this hypothesis, we used two independent methods, slot blot genomic hybridization and quantitative PCR (qPCR). Hybridization assays, while more material-intensive, have been used for many years; quantitative PCR is a more recent method

promising greater sensitivity and higher throughput and requires far less experimental material.

We evaluated the amount of five major heterochromatic repeats in each accession, relative to the Col-0 plant's genome, which we used as a comparison standard in all assays. The sequences assayed are the 180 base pair (bp) centromeric repeat (AR1), ORF1 of the high copy number pericentromeric Athila transposable element, fragments of the 18S and 25S ribosomal RNA genes, and the 5S RNA gene. The primers used to amplify these sequences are given in Materials and methods.

Measurements of the relative amount of the major heterochromatic repeats in the five accessions are presented in Table 2.2. We assayed one individual in each accession by both slot blot hybridization and quantitative PCR and assayed an additional individual, a sibling, in each accession using only qPCR. The data are graphed for each repeat assayed in Figure 2.3 against the measured genome size of the accession, relative to the Col-0 individual used as a standard. The assays reveal the presence of broad polymorphism in copy number of the repeats; the measured amounts of the centromeric repeat, the pericentromeric transposable element Athila, and 45S rDNA vary by over a factor of two, and the 5S rDNA cluster by a factor of four.

In Figure 2.3, panels A) and B) present qPCR results for the 18S and 25S ribosomal RNA genes; the 45S gene's RNA is cleaved following transcription into these two and the 5.8S ribosomal RNA. We assayed the two subunits separately to assess the utility of the method in our study. Because 18S and 25S are single-copy subunits of a larger RNA sequence, their RNV among accessions may be expected to be identical; linear regression between the two separate assays gives a coefficient of determination r^2=0.71 (p-value = 0.002), indicating good agreement. As noted we measured the copy number of each RNA gene in siblings (black and white markers in the figure) for each accession. For both subunits the difference in measured copy number between siblings is less than the standard error of the mean. One puzzling piece of our data is the measured

amount of these genes in two siblings (circled) of the Col-0 individual used as a comparison standard. These two sibs agree closely, as do the others, but differ from the standard accession by more than one standard error. Across all measured repeats the two have a mean value of 1.2 times the comparison standard.

We pooled the 18S and 25S RNA genes' probes in our slot blot measurements, these data are presented separately from the qPCR values in panel C). The remaining panels in this figure present the slot blot and qPCR results together: slot blot values for each accession are indicated by a white diamond; the qPCR result for the same individual is given by the white square. The black square presents the qPCR result for the sibling. Panel D) gives these data for the 5S ribosomal gene; we found this repeat exhibits more variation in copy number than any of the others assayed.

To test this finding we prepared FISH slides of anthers from the reference accession Col-0 and the accession with the highest measured 5S rDNA copy number, Loh-0. We probed Arabidopsis nuclei with a fluorescently labeled fragment of the 5S gene; representative images from the assays are given in Figure 2.4. The *in situ* results corroborate differences in amount of the 5S repeat in the two accessions. Panels (A) and (B) present images of meitotic pollen mother cells in the two accessions. The set of Loh-0 5S rDNA images was scored as significantly brighter than Col-0 (chi-squared p-value < 0.005) by four observers; 20 Col-0 and 22 Loh-0 cells were scored in each set. Panel (C) presents a second set of images, of diploid cells also from anther squashes. *In situ* hybridization identifies six 5S rDNA loci in both accessions; in Col-0 one locus each on chromosomes 3, 4, and 5 is known to be present, giving six in the diploid nucleus.

In the images the Loh-0 accession clearly exhibits a greater difference in intensities among the hybridized loci than Col-0. We quantified the intensity of the six 5S spots in 11 Col-0 and 16 Loh-0 images and chart the averages in panel (D). Because the loci are not individually identifiable we ranked them by

intensity. In the averages we find a stair step rather than the expected three pairs of equal homologs; variation in chromosome orientation on the slides and partly obscured sequences will result in signal diminution and produce this effect. Despite that the chart shows that the lowest intensities are nearly equal in the two accessions, and the highest is significantly brighter in Loh-0 than in Col-0 (ANOVA p-value = 0.03). From this we infer that a gain in the former accession, and not a loss in the latter, contributes to the greater contrast among the 5S hybridization spots in Loh-0.

Figure 2.3, panels E) and F) present results for the 180 bp centromeric repeat and the abundant pericentromeric transposable element Athila, respectively. While we find significant differences in both these repeats in the accessions, the accession with the largest measured genome size, Loh-0, has the same copy number or less of each as Col-0, whose measured genome size is approximately 12 Mbp smaller. This is true also of the measured copy number of 18S and 25S genes. The Loh-0 genome may differ from Col-0 principally in the possession of additional 5S array loci, or variation in size may be contributed by elements other than the tested repeats.

The qPCR assays identified larger differences in the other repeats between siblings than the average 10% in the 18S & 25S ribosomal RNA genes, with a mean difference of 24% in 5S rDNA, 21% in the 180 bp repeat, and 16% in Athila. While Arabidopsis is almost entirely a selfing plant and is expected to be homozygous, development of polymorphism in heterochromatin of inbred plants has been reported (Weimarck, 1975). Overall the measured differences between siblings are a small fraction of that determined among the accessions; over repeated generations, however, drift in the copy number of these elements could contribute to large differences.

Comparative Genomic Hybridization assays

To complement our assays of repeat polymorphism we sought a genome-wide assay of genic CNV. We wanted to determine, for example, if Loh-0 had one or

more segmental duplications of chromosomes that could contribute to size differences. We employed comparative genomic hybridization (CGH) microarrays to assay the copy number of genic sequences in the accession with the largest measured genome (Loh-0), compared with the sequenced Columbia accession (detailed in Materials and methods). After quality control of the hybridization data, some 18,000 hybridized features remained for this analysis. The microarray oligos are designed from known genes, EST sequences and predicted transcripts.

Figure 2.5 presents two views of the hybridization results. Values charted are the base 2 logarithm of Loh-0 feature intensities, relative to Columbia; see the caption for a detailed explanation. This oligo set cannot assay the copy number of intergenic sequences or the centromere cores as both are absent from the set; neither are the ribosomal RNA genes represented.

A number of transposable elements, a class chiefly closely associated with centromeres and nearby sequences, are present on the array. In the most recent (TIGR5) Arabidopsis genome there are 519 pericentromeric Athila genes, alone totaling 1.6 Mbp; Athila is known as well to be a major component of the centromere cores. In our CGH dataset there are 190 known transposon-related features available; thus this class of genes is deficient in our array data, especially on chromosomes 4 and 5, relative to their known presence in pericentromeric regions of the genome. Feature density in the chromosome arms is 1 per 6,000 bp and in the pericentromeric regions 1 per 14,000 bp; on chromosome 5 the near-centromere value is 1 per 30,000 bp.

Regional differences are apparent in these data. The pericentromeric features of chromosomes 1, 2 and 3 have the lowest Loh-0/Col-0 ratios of any comparable size regions in the genome. The pericentromeric region's ratios, as defined by the presence of Athila elements in the TIGR sequence, have a mean of 0.97. The same value for the chromosome arms is 1.02. It is clear from the CGH array that no large-scale segmental duplication of genes is available to account for the extreme measured genome size of the Loh-0 accession; the

near-centromere transposable element ratios support the qPCR results of the identity or small reduction in this genomic class amount relative to the sequenced genome Col-0. Note that the normalization applied to the ratios in order to correct for microarray block and dye intensity-dependent effects constrains the global mean to exactly one.

Modeled contribution of the repeats to genome size variation

We developed a simple data model to assess whether our measured repeat fractions could account for the different genome sizes we determined with flow cytometry. The model first calculates the size of the variable genome in each individual as the size in Mbp of each of the heterochromatic elements in the sequenced genome times the individual's qPCR-measured repeat amount relative to Col-0; the combined size of the sequenced genic and intergenic regions (108 Mbp) is added to give the total genome size. As the sequenced genome's heterochromatin repeat sizes are not known with precision, the model tests a series of sizes for each repeat, drawing on published size estimates to establish a range. We designed a merit function (Press *et al.*, 1986) to assess agreement between the flow cytometry-measured and model-predicted genome sizes and used it to identify Col-0 repeat sizes giving the best overall fit. Because of the unsettled understanding of the size of the Arabidopsis genome, we determined separate sets of these values for Arabidopsis genome sizes of 130, 145, and 160 Mbp; the model is described in Materials and methods. We used only the qPCR results in this analysis.

We found variation in the four large repeat arrays we assayed accounted for up to 61 percent of the variance in measured genome size among the accessions, when we included all assayed accessions in the analysis. We tested the model both with and without the accession with the largest measured genome, Loh-0, thinking that its extreme genome size and pattern of variation could challenge the model. When omitting this accession from the model the assayed differences in four repeats explained up to 49 percent of measured

genome size variation. In the former case the model analyzed ten individuals in five accessions, and eight in four accessions in the latter. Figure 2.6 presents these results.

As we outlined, the model tests the linear relationship Y = mX + b where Y is the flow-measured genome size of an accession, b is the basal (genic and intergenic) genome size of 108 Mbp, and X is the numerical value of repeats in an individual — our model shuffles the possible values of Col-0 repeat sizes to best fit the equation across all individuals. The constant (m) is a scaling factor that extends the model's ability to fit both sides of the equation by stretching or shrinking the right side. A single value of (m) applies to all the individuals considered, and differs from m = 1 only if there is distortion in the measurements — a scaling error. We found for all runs of the model, a value of m = 0.8 produced the best fit. This result may resolve our finding that the Col-0 individuals we assayed as checks on the accuracy of repeat measurements consistently returned values one and a quarter times the sibling's used as a standard; the 0.8 scaling factor lowers these ratios to unity.

Why the test siblings should differ from the standard in this way and the sources of this scaling value are not clear. Several explanations are possible; two relate to our tools and their interaction with the material. The qPCR method we used may exaggerate copy number differences; the checks we employed in each assay with known differences in copy number reveal the method does often underestimate or overestimate them, but not systematically. Another possibility is that flow cytometry under-reports genome size differences. This could occur, if for example, the larger genomes' chromatin is systematically less open than the smaller, or if the DNA of the larger is under-replicated with respect to the smaller.

Discussion

Our flow cytometry measurements document significant differences in measured genome size (MGS) among accessions of *Arabidopsis thaliana*. The accessions are from geographically separate populations, and Arabidopsis is largely (98%) a

selfing species. We found the differences in genome size among individuals in a single accession are far less than that found among accessions. Our findings agree with those reported by Schmuths *et al.* (2004), that the sequenced accession Columbia has one of the smallest measured genomes, and that the maximum difference among accessions is around 10% of the mean. We measured the GS of Columbia to be 157 Mbp; that report's value is 202 Mbp.

Inter-laboratory comparison of plant genome size estimates by flow cytometry (Dolezel *et al.*, 1998) documented the absence of agreement in absolute value among measurements made in independent laboratories, even using the same material for samples and standards. The maximum difference for the genome size of Arabidopsis accession Columbia among the four participating laboratories was 42%; significantly, intra-laboratory variation in duplicate measurements was reported as negligible. The reasons for disagreement are not well understood; factors including use of different instruments have been noted. It is known that propidium iodide (PI) fluorescence varies with chromatin condensation — the DNA double helix must unwind to permit intercalation of the molecule. Darzynkiewicz *et al.* (1984) found that rapidly growing leukemia cells stained with PI gave a (normalized) fluorescence of 1.00; differentiated cells' fluorescence was 0.94. After extraction of nuclear proteins with 100 mM HCl, measured fluorescence of undifferentiated cells was 2.03. We found that fixing Arabidopsis nuclei in ethanol, a process known to modify chromatin condensation, increased measured genome size by 12 percent. Thus, fluorescent staining of nuclei can be affected by factors other than genome size.

We identified large-scale polymorphism of heterochromatic repeats in Arabidopsis accessions with two independent methods; they gave similar but not identical results. The correlation coefficient of the slot blot hybridization 45S RNA copy number with the averaged qPCR 18S and 25S values is 0.60, and a linear regression between the two sets of data gives a p-value of 0.28 — one is not a good predictor of the other. For all assayed repeats the correlation between the slot blot and qPCR measurements averages 0.66; it is weak for the centromeric

repeat (r = 0.29), moderate for the 45S RNA (0.60), and strong for the Athila retrotransposon and the 5S RNA repeats (0.81 and 0.96 respectively). This variation may be due to differences in what each method measures: slot blot data reflect hybridization of probes up to 700 bp with genomic sequences fixed on a membrane; in qPCR genomic sequences around 125 bp are amplified through multiple cycles, directed by hybridized 20 bp primers. Because qPCR assays depend on the conservation of its primers, SNPs in or deletion of these short sequences will result in errors in template estimates. Differences in sequence conservation between the 5S RNA gene and the centromeric 180 bp arrays, for example, may contribute to the differences we found in agreement of the two methods for these repeats.

We were particularly interested to determine whether variation in a single repeat could explain genome size differences among accessions. The strongest positive relationship between them in our data is in the 5S RNA array; the correlation with measured genome size with all our 5S data is 0.80, and a linear regression between the two returns a p-value of 0.0003, highly significant. The previously estimated size of this repeat in Col-0, 1 Mbp, however, is small compared with the measured genome size variation. The correlation between genome size and the 45S RNA array findings is weak (r = 0.34).

Surprisingly, the correlation between genome size and the Athila repeat is negative, r = -0.53 for all measurements, with a significant linear relationship (p-value = 0.04). Most surprisingly to us the correlation between genome size and the 180 bp centromeric repeat is negative, a weak -0.14 in the qPCR data, but a remarkable -1.00 in the slot blot results. The linear regression between genome size and this repeat for all data is not significant (p-value = 0.51), but highly so in the slot blot data taken alone (p-value = 2×10^{-5}). As the 180 bp and Athila repeats are grouped together and intermixed, this finding is particularly intriguing.

We are able to assign size values to the heterochromatic repeat arrays using a computational model that takes into account relative measured genome sizes of the accessions and repeat copy number measurements. The model's

best fitting value for the size of the Arabidopsis genome, 145 Mbp, is close to recent estimates, but less than our measured 157 Mbp using flow cytometry. The centromeric repeat sizes allocated to the genome by the model are larger than the AGI (2000) figure, but only with a 160 Mbp Arabidopsis genome do they approach the later assessments in Table 1.1. The model assigns additional base pairs to the 45S rDNA arrays when Loh-0 is excluded from its consideration, with the 5S rDNA arrays given a value near that estimated in the literature. With that accession and its extreme 5S rDNA measurements included, the model shifts base pairs to those repeats. The model-determined Athila TE values are close to the sequenced amount of 1.6 Mbp.

In our data model, measured differences in the four assayed repeats account for up to 60% of the flow cytometry-determined genome size variation at best. This result may imply that variation among the unassayed repeats in the genome plays a significant role. Over 200 repeat families have been identified in the sequenced genome of Arabidopsis, with 28,000 individual elements totaling at least 12 Mbp (Jurka, 2000); their intraspecific variation has not been assessed, to our knowledge. Alternatively, the negative relationship we found between number of heterochromatic repeats in and near the centromeres and the flow cytometry-measured genome size, suggests the possibility that differences in pericentromeric chromatin compaction among accessions modulate intercalation of the fluorescent dye, affecting accuracy of the method. Contrary to our expectations no single repeat array's variation explains the differences in genome size among the accessions, nor did we find a simple pattern of repeat polymorphism. Positive and negative differences in all four of the repeats result in net genome size differences.

Microarray comparative genomic hybridization assays complement this finding in showing that no comparable large-scale continuous duplication of genes or chromosomal rearrangement is present in the largest measured genome, relative to Columbia. The inset in Figure 2.5 demonstrates the method can readily differentiate between two and three chromosome copies. Deletions

on the order of 100 Kbp in Loh-0 are indicated, for example at 8.8 Mbp and 21.3 Mbp on either side of the chromosome 1 centromere. Insertions with respect to the Col-0 reference can be detected only if they are represented in the array: thus, hypothetical sequences that are unique to Loh-0 are not examined.

The significance of heterochromatin polymorphism within a species either to the individual or to evolutionary mechanisms is unclear, but methods available today could shed light on fundamental questions picked up earlier and set aside as intractable. For example, so far as we know the Arabidopsis accessions with heterochromatin copy number differences are stable lines. Hybridization between them may further our understanding of the processes shaping these genomic regions and their interaction with euchromatin.

Materials and methods

Arabidopsis accessions and growth conditions

We acquired Arabidopsis accession seed from the Arabidopsis Biological Resource Center (ABRC) at Ohio State University, and from Prof. Magnus Nordborg at the University of Southern California. The ABRC stock number and identification provided by Prof. Nordborg are listed in Table 2.1 with the accession name. Seed were sown directly on wet potting soil in 2 inch pots, maintained in the dark at 4°C for four days, and moved to a 22°C growth room where the plants germinated and were grown under fluorescent lights with 16 hours light and 8 hours dark per day.

Genome size determination with flow cytometry

Genome size measurements were made at the Cell Analysis Facility of the Department of Immunology, University of Washington; a Becton-Dickinson FACScan flow cytometer with 488 nm argon laser was used. Linearity of instrument response to DNA content was assayed using aggregated chicken erythrocyte nuclei.

Sample preparation was as follows. Stained nuclei: 100-300 mg of leaves were collected and stored temporarily in a petri dish on ice. Chopping buffer (1.5 ml) was added to the dish, and leaves chopped with a razor blade, mixing until a paste, 2 to 4 minutes. Liquid was collected and aspirated with a syringe; filter holder (Millipore Swinnex 25 mm) attached with 30 µm filter fitted inside (Small Parts Inc CMN-30 monofilament cloth), and pressed through the filter into a microfuge tube. Tubes were spun at 500Xg for 7 minutes; supernatant discarded and 3 µl of the internal standard added, chicken erythrocyte nuclei (Becton-Dickinson DNA QC particles, Cat. No. 349523, or BioSure chicken erythrocyte nuclei singlets, Cat. No. 1013), and nuclei resuspended in 700 µl staining solution. Samples were capped and stored above ice at least 2 hours prior to evaluating DNA content, and protected from light. Chopping buffer: modified from Bino *et al.* (1993), 15 mM HEPES, 1 mM EDTA, 80 mM KCl, 20 mM NaCl, 300 mM sucrose, 0.20% Triton-X, 0.5 mM spermine, 0.10% β–mercaptoethanol (BME). Buffer without BME may be stored at 4°C indefinitely; BME is added just before use. Staining buffer: 50 µg/ml of the fluorochrome propidium iodide (PI) and 50 µg/ml RNAse A was added to chopping buffer. PI is a potential mutagen and handled accordingly.

Figure 2.7 illustrates the method. Panel (A) presents a typical result. The position of the chicken erythrocyte nuclei relative to the Arabidopsis peaks permits a determination of the Arabidopsis accessions' genome sizes. Like many plants the Arabidopsis genome is endoreplicated in some of its tissues; in leaves diploid (2C) nuclei as well as tetraploid (4C) and octaploid (8C) nuclei, and higher ploidies are present. In panel (A) the higher ploidy nuclei are masked or off-scale. Panel (B) shows the relative intensities of nuclei from two accessions (Ta-0 and Loh-0) with measured genome sizes differing by 20 Mbp (see Figure 2.1). Both samples were prepared together for this assay; more Ta-0 than Loh-0 cells were present in the material, giving a larger peak.

DNA extraction

Plant DNA was extracted from 1 gm rosette leaves, ground for several minutes in a mortar, initially with a small amount of liquid nitrogen to facilitate reducing the leaves to powder. Plant extraction buffer (150 mM Tris pH 8.0, 50 mM EDTA, 500 mM NaCl, 0.7% SDS, 50 µg/ml Proteinase K, 50 µg/ml DNAse-free RNAse A) was added to a total volume of 8 ml during grinding. The sample was filtered through Miracloth and heated in round-bottom tubes in a water bath at 55°C for 3-5 hours; 4 ml saturated NaCl was mixed in each tube and spun in a preparatory centrifuge at 7,000Xg for 20 minutes. The supernatant was divided into 2 tubes and 7 ml 85% isopropanol added and mixed by inverting; supernatant was discarded after spinning again for 10 minutes, the pellet washed twice in 70% ethanol, and air-dried for 10 minutes. The pellet was resuspended in 1 ml TE and transferred to a 1.5 ml tube; 1 µl 25 mg/ml RNAse A added and incubated at 37°C for one hour. The procedure was completed with phenol extraction and ethanol precipitation and washing, and after air-drying the sample was resuspended in TE and frozen at -20°C.

Slot blot hybridization

Biodyne nylon transfer membranes were cut to fit in the Bio-Rad Bio-Dot SF blotting apparatus with 48 wells; each well is 7x0.75 mm. Membranes were loaded with genomic DNA extracted from 2 individuals, one the single standard loaded on each membrane, the second a test plant; before loading on blots the DNA was needle fragmented. DNA concentration was quantified using a Turner fluorometer with SYBR green dye from Molecular Probes and a lambda-phage DNA standard; when it became available sample DNA concentration was re-assayed with a Perkin Elmer Victor3 V plate reader. Before loading, DNA extracts were heated to 100°C in boiling water for 10 minutes, immediately cooled on ice, and diluted to 1 ng/µl in 0.4 M NaOH. Each sample was loaded in 8 slots in one of 3 amounts, 100, 125 or 150 ng for a total of 24 slots per plant

distributed across the array to assay linearity of fluorescence with hybridization. The loaded DNA was neutralized by floating the membrane on 100 mM Tris pH 8, cross-linked to the membrane with a UV Stratalinker and allowed to air dry before use.

The Amersham Biosciences AlkPhos Direct Labeling/Enhanced Chemifluorescence System was used to fluorescently label the DNA probes, hybridize probes to membrane-bound genomic DNA, and develop the hybridized labeled probe according to the manufacturer's instructions. Fluorescence was excited and detected with the UVP Epichemi3 Darkroom/Benchtop UV Transilluminator with filter set to 515-570 nm, and blot images captured with a digital camera. Signal intensity was quantified with the ImageJ open source gel blot analysis software available from the Research Services Branch of the U.S. National Institutes of Health. Blots were stripped of hybridized probe according to the manufacturer's instructions, stored in 100 mM Tris pH 8 at 4°C and reused.

DNA probes from 120 to 700 bp in length were generated using the PCR of DNA extracted from Arabidopsis accession Columbia-0 with the following primers: the 180 bp centromeric repeat (5'-CAT GGT GTA GCC AAA GTC CAT A-3' and 5'-GCT TTG AGA AGC AAG AAG AAG G-3'; ORF1 of the Athila retrotransposon was amplified using degenerate primers and a touchdown thermocycler program as described in Josefsson *et al.* (2006). The 5S rDNA gene primers were (5'-GAT GCG ATC ATA CCA GCA CT-3' and 5'-GGA TGC AAC ACG AGG ACT TC-3'), 18S rDNA gene (5'-GCA TTT GCC AAG GAT GTT TT-3' and 5'-GTA CAA AGG GCA GGG ACG TA-3'), and 25S rDNA gene (5'-AGA ACC CAC AAA GGG TGT TG-3' and 5'-TCC CTT GCC TAC ATT GTT CC-3').

The amount of heterochromatic repeat in each accession relative to the single standard was calculated as the ratio of the accession's mean value (A) on a blot divided by the standard's (B). To estimate uncertainty in the results, the standard error of each measured value (ΔA, ΔB) was used; the relationship of the final parameter to the measured variables was used to propagate standard

errors. For a function of two variables the uncertainty $\Delta F(x,y) = [(\partial F/\partial x)^2 (\Delta x)^2 + (\partial F/\partial y)^2 (\Delta y)^2]^{1/2}$. In the slot blots with F(A,B) = A/B, the fractional standard error $\Delta F/F = [(\Delta A/A)^2 + (\Delta B/B)^2]^{1/2}$.

Correct assessment of relative amounts of a sequence using blot hybridization requires accurate determination of DNA concentration in the target DNA extracts deposited on the membrane. We measured DNA concentration using fluorometric analysis of SYBR green stained DNA and a 96-well plate reader.

We found that variation on a membrane among hybridization signals of the same probe and target from slot to slot was sufficiently large to require multiple replicates of a target. We designed a loading pattern to minimize the effects of spatial heterogeneity in hybridization, and also to enable estimation of the linearity of hybridization signal with target amount. The design, shown in Figure 2.8, loads one test sample along with the comparison standard used on all membranes. A representative blot image is also presented in this figure. In analysis, measurements of signal from fluorescently-labeled probe hybridization to a sample's 24 slots were averaged, along with the standard's, and the ratio of the two (test/standard) was taken as the amount of probe sequence in the test sample, relative to the standard.

Quantitative PCR

Quantitative PCR reactions were run in 96-well plates in a Chromo4 Continuous Fluorescence Detector and Thermocycler from MJ Research, Inc; initial data analysis was made using the Opticon Monitor software from the same company. The individual DNA samples used in the slot blot assays were also used in these assays. Replicates (from 6 to 12 of each sample and amplicon) were loaded distributed across a plate, using DNA in 3 amounts, 1.00, 1.25 and 1.50 times a basal loading, in order to assess linearity of amplification and detection. DNA extracts used in qPCR reactions were diluted 25X in water before use; the fluorophore used was SYBR green.

Reaction volumes were 20 µl with the following reagents (per reaction: 11 µl water, 1.6 µl 2.5-mM dNTPs, 0.2 µl 20-µM primers, 0.05 µl 100X-SYBR green, 0.2 µl 5U/µl-*Taq* polymerase, 2 µl 10X-buffer, and 5 µl genomic DNA, approximately 5 ng/µl). The thermocycler protocol was (94° for 120 seconds, then cycle 40 times: 94° for 20 seconds, 57° for 20 seconds, 72° for 30 seconds; using a heated lid at 100°). A melting curve was generated for each reaction product to test for multiple amplification products. Amplicon template quantities were measured using a threshold cycle (C_t) method (Peirson *et al.*, 2003; Bubner *et al.*, 2004). See Larionov *et al.* (2005) for a discussion of error analysis and reduction. Briefly, for each amplicon, a dsDNA fluorescence value was selected where all accessions' templates had been amplified to the same copy number; the threshold cycle where this occurred was recorded for each accession. Relative amounts of initial template quantity Q_o were calculated with the relationship $Q_o = A^{-Ct}$ where A is the cycle amplification factor. Replicates were averaged to provide a single value for analysis. Estimates of uncertainty used the standard error of the individual estimates of A and C_t with the errors propagated as in the slot blots. The propagated fractional standard error $\Delta Q_o/Q_o = [(C_t/A)^2 (\Delta A)^2 + (\ln(A))^2 (\Delta C_t)^2]^{1/2}$.

The copy number of the assayed repeats in each sample's genome was measured as the ratio of template amount of the repeat to the template amount of single copy gene amplicons. Amplification products were from 130 to 300 bp long. The following primers were used: for the 180 bp centromeric repeat (5'-CCG TAT GAG TCT TTG GCT TTG-3' and 5'-TTG GTT AGT GTT TTG GAG TCG-3'); probes of the retroelement Athila were derived from *A. thaliana* sequences amplified with degenerate primers and cloned (Josefsson *et al.*, 2006). Representative clones were aligned to identify conserved sequences from which these primers were designed; Athila ORF1 (5'-TTT CTC ACT AGG GGA TAA AGC TCA-3' and 5'-CAA TCT AGC CGT TCT TGA GTT AGA-3'). Primers for the 5S rDNA gene were as in the slot blot hybridization; for the18S rDNA gene (5'-CCT GCG GCT TAA TTT GAC TC-3' and 5'-GAC AAA TCG CTC CAC

CAA CT-3'), and 25S rDNA gene (5'-CGC GAG TTC TAT CGG GTA AA-3' and 5'-CAC TTG GAG CTC TCG ATT CC-3'). Single copy genes used were actin (ACT2, At3g18780) (5'-TGC CAA TCT ACG AGG GTT TC-3' and 5'-TTA CAA TTT CCC GCT CTG CT-3), and cyclophilin (ROC1, At4g38740) (5'-TCA AGC CAA TCG GTC TTC AC-3' and 5'-CGA TCT ACG GGA GCA AGT TC-3').

We assayed DNA extract genome copy number using qPCR of two single copy genes and independently confirmed those results with excellent agreement ($r = 0.97$, data not shown) using a plate reader and the fluorescent dye SYBR green to measure DNA concentration.

Quantitative PCR has the potential to accurately assay amplicon template quantity of single copy as well as repeat sequences in genomic DNA with a sensitivity not previously achievable. Processing its measurements entails greater analytical complexity than slot blot hybridization. The method is based on the detection of an amplicon's exponentially increasing amount at each cycle of the polymerase chain reaction. Estimates of initial template quantity are derived from its amplification rate and quantity after a known number of cycles.

We found as with slot blot hybridization that multiple replicates of each reaction were required to limit uncertainty in our results; we ran either 6 or 12 replicates well-distributed on a 96-well plate, giving either 16 or 8 unique reactions per plate. Previous work has shown that qPCR can assay differences in template amount less than 25%, using purified amplified DNA sequences as templates (Gentle *et al.*, 2001). A separate study again using a PCR product achieved precision between 6% and 21%, precision falling with increasing threshold cycle C_t (Rutledge & Côté, 2003). Comparison of genomic DNA is potentially more challenging.

Table 2.3 presents representative values of observations with replicates of a single amplicon in several accessions, with values for amplification factors and threshold cycles; discussed more fully in Materials and methods. To determine estimates of genome copy number in our DNA extracts, we used qPCR measurements of the amplicon template amount of single-copy genes. These

genes occur only once in the sequenced accession Col-0. Two genes, actin (ACT2, At3g18780) on chromosome 3N, and cyclophilin (ROC1, At4g38740) on chromosome 4S, were used for this purpose. It is possible these genes are themselves present in different numbers of copies in the accessions assayed. Linear regression between copy number of both genes measured in all accessions gave a coefficient of determination (r^2) of 0.92, from which we infer as they are unlinked both genes are most likely single-copy in the accessions.

Fluorescent In Situ Hybridization (FISH)

The plant material was prepared as in Comai *et al.* (2003), with the following changes: the *A. thaliana* 180 bp centromeric repeat probe fluorescent dye was fluorescein-12-dUTP (FITC, Roche 1373242) and the 5S rDNA array probe fluorescent dye was tetramethyl-rhodamine-5-dUTP (Roche 1534378). Probes were amplified with the primers identified in the slot blot method. Prepared slides were visualized using the Nikon Microphot-FX fluorescent microscope; images were captured with the Qimaging Retiga 1300 monochrome 10-bit digital CCD camera and processed with the Improvision Openlab image analysis software, V4.0.4. Camera exposure times were chosen to maximize image clarity without saturation of pixels. Photographs were taken, and reviewed and cropped using Adobe Photoshop; no additional image enhancement was performed.

To assess relative amounts of the 5S rDNA repeat in the Col-0 and Loh-0 pollen mother cells, anther squashes of both accessions were prepared side-by-side on slides and probed. In the scoring process, 20 Col-0 and 22 Loh-0 images were randomly presented in grayscale using the JPEGDeux open source slideshow application. The scoring individual identified the 5S intensity in each image as either plus or minus without knowing the accession (blind scoring). Four people independently scored the set of 42 images, and all identified the Loh-0 accessions as significantly brighter (chi-squared p-value < 0.005). Combined scores for each accession are 5% plus for Col-0 (plus and minus = 4 and 76), and 66% plus for Loh-0 (plus and minus = 59 and 29).

For assessment of individual 5S rDNA locus hybridization intensity in the diploid cells, Col-0 and Loh-0 anther squashes were prepared as above, on separate slides and treated to clear the nuclei after slide preparation with 2 µg/ml pepsin A (Sigma P6887-250MG) in 10 Mm HCl for 10 minutes at 37° in an incubator, and washed twice for 5 minutes each in 2X SSC. Pepsin was not used with the pollen mother cells in the previous assay; additional washes increase the risk of intermixing the two accessions' cells. The open source Java application ImageJ was used to quantify 5S locus size in each image. Image background was first subtracted. Each 5S spot intensity was then measured as the mean pixel intensity in a circle centered on the spot. In the analysis spreadsheet the six spots in each cell were sorted by their intensities. We then calculated the mean and its standard error for each of the six spots over all measured cells.

Comparative genomic hybridization

Ratios measuring the relative amount of 26,090 70-mer sequences in two accessions' genomes were derived in the following way. For each of the two samples to be combined and assayed, 300 ng of unfragmented genomic DNA were labeled with either Cy3-dUTP or Cy5-dUTP (Amersham Cat. PA53022 or PA55022) using the Invitrogen Bioprime Array CGH Genomic Labeling System Cat. 18095-12, according to the manufacturer's instructions.

The Cy3 and Cy5-labeled samples of each accession were then combined and purified using the Qiagen QIAquick PCR Purification Kit Cat. 28104, according to the manufacturer's instructions. Yeast tRNA (Invitrogen Cat. 15401-029) was added to the labeled DNA at a final concentration of 0.5 mg/ml and SSC at 3X final concentration in 120 µl total volume. Each sample pair was also dye-swap labeled.

The labeled DNA samples were hybridized to a spotted microarray of the Operon Arabidopsis Genome Oligo Set Version 1.0 and washed and scanned at the DNA Array Facility of the Fred Hutchinson Cancer Research Center. Image conversion was done with GenePix Pro 6.0 and these data analyzed with the

TIGR open source Microarray Data Analysis System, MIDAS (Saeed *et al.*, 2003). Dye-swap pairs were filtered to discard features with signal-to-noise ratio less than two, and LOWESS normalized separately before being combined. Dye-swap consistency was checked, integrated feature intensities of each channel were written, and ratios of relative intensity calculated.

Further analysis of the data was carried out using the open source application CGH-Explorer, available from the Department of Informatics, University of Oslo (Lingjærde *et al.*, 2005), and the Microsoft Excel spreadsheet application.

Modeling heterochromatin contributions to genome size

We developed a numerical data model to provide an estimate of the absolute contribution of each of the heterochromatic repeats to the sequenced Arabidopsis genome. The model minimizes the difference between the set of flow cytometry-determined genome sizes and the set of genome sizes calculated from the repeat sizes measured by qPCR plus a basal, constant genome component. We estimated the last element at 108 Mbp, taking the sequenced amount of 115 Mbp (AGI, 2000), subtracting a sequenced 5 Mbp reported reported there from the centromeres, and 1 Mbp apiece for the Athila TE and 5S rDNA repeats. The 45S rDNA arrays were not sequenced. The modeled relationship between the two sets is $\mathbf{Y} = m\mathbf{X} + b$ where the symbol meanings are:

$\mathbf{Y}$	vector of measured genome sizes, with element y_i for the i^{th} individual (Mbp)
$\mathbf{X}$	vector of the summed repeat sizes with element x_i for the i^{th} individual (Mbp)
c_j	size of the j^{th} repeat in the sequenced Col-0 accession (Mbp)
w_{ij}	fractional amount of the j^{th} repeat in the i^{th} individual, relative to Col-0
x_i	repeat total in an individual: $x_i = c_j w_{ij}$ with j summed over all repeats (Mbp)
m	a scaling factor between repeat size and genome size measurements
b	the basal, unvarying genome component (108 Mbp)

The relationship is a simple linear one: we expect the amount of polymorphic repeats and the basal component to add up to the measured genome size of an individual. The scaling factor (m) is present to correct for any linear distortion in the response of the qPCR system to difference in repeat amount — if for example a 30% difference is measured as 40%. Ideally m = 1, but is not be assumed to be.

The computational model is written in Perl; it first reads for each assayed individual its ID, the measured genome size of the accession and the sizes of the centromeric, Athila, and 45S and 5S RNA arrays relative to the comparison standard Col-0 individual. We use the accession mean rather than the individual measured genome size because separate flow cytometry measurements of any individual appear to be randomly distributed around the accession mean; we average the 18S and 25S qPCR repeat size values to form a single 45S measurement.

Numerical values for repeats in Mbp are calculated for each individual by summing the products of the fractional size of each repeat in that individual (relative to the Col-0 standard) times the size of the repeat in the sequenced accession Col-0. The latter values are not precisely known as the repeats are unsequenced. The model sequentially assigns values to each repeat in Col-0 from a range of potential sizes. Most combinations of four repeat and basal genome size do not sum to the assigned Col-0 genome size and are discarded — the Col-0 total genome size is specified as a particular value for each model run. A merit function for each of the combinations passing this screen is calculated in the following way: the genome size of each individual is summed from its component parts; the standard deviation of the difference between the measured and calculated genome sizes (the RMS error) is then divided by the correlation between the two sets of values. If the correlation is less than 0.1 the combination is discarded. Optimal values of the scaling factor (m) are also assayed. The Perl script writes out the merit function value, associated repeat

and basal genome sizes and scaling factor to a file and proceeds with the next combination. Combinations with the smallest merit function are reviewed.

Statistical analyses

Statistical tests were evaluated using Microsoft Excel and its data analysis tools. The p-value reported for linear regressions is the regression tool ANOVA table's F-test results (Significance F). The 5S rDNA chi-squared test results were assessed using a table of critical values for the chi-squared distribution; the ANOVA p-value is that calculated by Excel's single factor ANOVA tool.

Table 2.1. Survey of genome size variation in *A. thaliana* determined by flow cytometry. Units are millions of base pairs (Mbp). Accession identification code is either the ABRC stock number (those starting with CS) or that given by the laboratory of Magnus Nordborg. CV: Coefficient of variation is the standard deviation presented as percent of the mean. Prepared samples of each individual were measured two to three times.

Average genome size (Mbp)	Size relative to Col-0	Accession name	Identification code	Standard deviation	CV	No. of individuals measured
149.3	0.95	Ta-0	CS1548	0.8	0.6	2
153.1	0.98	Lip-0	CS1336	0.2	0.1	2
154.6	0.99	Br-0	9A Br-0 A	2.4	1.5	2
155.2	0.99	Sp-0	CS1530	2.6	1.7	2
156.5	1.00	Col-0	8F Col-0 A	0.8	0.5	5
157.5	1.01	GOT-7	6E GOT-7 B	2.2	1.4	2
158.0	1.01	GOT-22	6F GOT-22 A	0.5	0.3	2
158.3	1.01	Rsch-0	CS1490	3.6	2.3	2
158.4	1.01	Ag-0	9C Ag-0 A	0.2	0.2	2
159.0	1.02	Yo-0	8E Yo-0 A			1
159.9	1.02	Nc-1	CS1388	3.6	2.2	2
160.0	1.02	Is-0	CS1240	3.0	1.9	4
160.5	1.03	Ct-1	CS6674	1.2	0.7	2
160.5	1.03	Per-1	CS1444	2.0	1.2	2
161.2	1.03	RRS-10	1B RRS-10 A	0.3	0.2	2
161.4	1.03	Tsu-0	CS1564	0.7	0.4	2
162.9	1.04	Eden-1	2A Eden-1 C			1
164.5	1.05	TAMM-2	6A TAMM-2 A	0.8	0.5	2
166.9	1.07	Kondara	11H Kondara A	0.4	0.2	1
169.1	1.08	Nok-3	CS6810	3.7	2.2	2
169.7	1.08	Wt-5	CS6896			1
169.9	1.09	Loh-0	CS1350	3.4	2.0	2

Table 2.2. Measured size of heterochromatic repeat arrays in five *A. thaliana* accessions. Units are based on the Col-0 standard (Col-0 = 1). Repeats in individual (A) were assayed using both slot blot genomic hybridization and qPCR, its sibling (B) was measured with qPCR only. Slot blot values are presented in regular font, **qPCR in bold**. In the slot blots probes for the 18S and 25S subunits of the 45S rDNA gene were pooled; the same value is presented for each subunit. SE: Propagated standard error of the mean, presented as percent of mean value. For measurement and error propagation details see Materials and methods.

Accession	Individual	18S	SE	25S	SE	5S	SE	CEN	SE	Athila	SE
Ta-0	A	0.84	9	0.84	9	0.68	6	0.90	8	1.27	10
Ta-0	A	**0.60**	**12**	**0.87**	**12**	**1.07**	**5**	**1.29**	**4**	**2.61**	**11**
Ta-0	B	**0.79**	**11**	**0.70**	**19**	**0.62**	**16**	**1.27**	**11**	**2.61**	**13**
Br-0	A	0.85	6	0.85	6	1.15	10	0.84	7	1.05	13
Br-0	A	**1.09**	**10**	**1.18**	**15**	**1.75**	**4**	**1.10**	**9**	**1.40**	**11**
Br-0	B	**1.04**	**8**	**1.13**	**13**	**1.71**	**6**	**1.67**	**11**	**2.12**	**15**
Is-0	A	1.61	12	1.61	12	1.75	7	0.78	5	1.54	11
Is-0	A	**1.32**	**16**	**1.63**	**21**	**1.82**	**4**	**0.87**	**8**	**2.57**	**12**
Is-0	B	**1.38**	**10**	**1.50**	**13**	**1.40**	**8**	**0.95**	**12**	**2.17**	**14**
TAMM-2	A	0.59	6	0.59	6	0.89	8	0.76	8	0.92	9
TAMM-2	A	**0.71**	**11**	**1.02**	**8**	**1.15**	**4**	**1.60**	**8**	**1.40**	**11**
TAMM-2	B	**0.77**	**12**	**0.90**	**20**	**0.84**	**20**	**2.21**	**10**	**1.65**	**13**
Loh-0	A	1.68	6	1.68	6	2.28	9	0.54	5	1.10	15
Loh-0	A	**0.99**	**9**	**1.00**	**8**	**2.61**	**4**	**0.98**	**9**	**1.04**	**12**
Loh-0	B	**1.04**	**10**	**0.90**	**20**	**2.82**	**9**	**1.26**	**11**	**1.10**	**14**

Table 2.3. Sample data from quantitative PCR assay of template copy number. Twelve replicates of one individual's extracted DNA loaded in three concentrations were used to amplify an exon fragment of the ACT2 gene (At3g18780); assays were made in 96-well plates. The estimated amplification factor (A) developed in each reactions, and the threshold cycle (C_t) where all reactions realized identical fluorescence, are given. The base 10 logarithm of the relative initial template copy number $Q_o = A^{-Ct}$ is calculated using the overall average amplification factor of the twelve reactions and the average threshold cycle for each concentration. Ratios of the measured initial template copy number of the higher concentration samples relative to the lowest were calculated as a check on agreement within the set of measurements.

Well	Relative template input	Amplification factor A	Ct	$\underline{A}$	$\underline{Ct}$	Log_{10} (Qo) by overall $\underline{A}$	Ratios of Qo's
C1	1.00	1.85	21.6				
C4	1.00	1.70	21.0				
F9	1.00	1.63	20.6				
F12	1.00	1.55	20.3	1.68	20.9	-4.9494	1.00
C2	1.25	1.74	20.9				
C5	1.25	1.78	20.4				
F8	1.25	1.70	20.3				
F11	1.25	1.79	20.3	1.75	20.5	-4.8580	1.23
C3	1.50	1.90	20.6				
C6	1.50	1.81	20.2				
F7	1.50	1.64	20.0				
F10	1.50	1.63	19.9	1.75	20.2	-4.7851	1.46
	Overall $\underline{A}$	1.73					

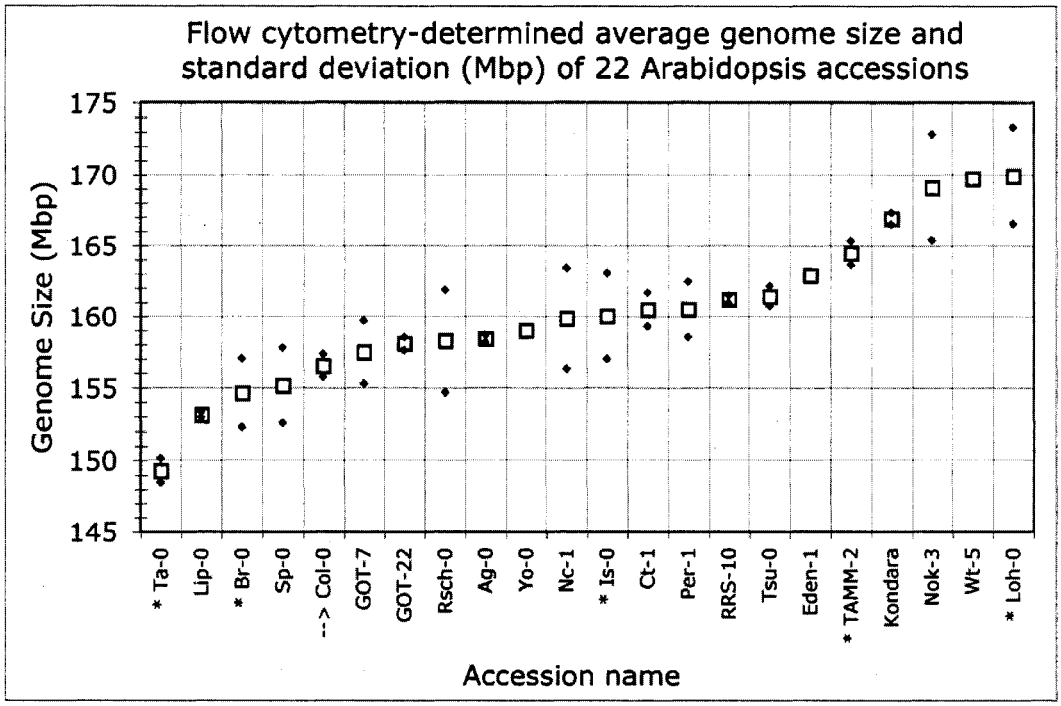


Figure 2.1. Mean genome size in millions of base pairs (Mbp) of 22 *Arabidopsis thaliana* accessions, measured using flow cytometry. Values are shown by a square marker; where more than one individual was measured the standard deviation of the separate means is given by diamond markers. Accessions selected for more detailed genome size measurement and assessment of repeat array size are marked with asterisks; the sequenced accession Col-0 is noted with an arrow.

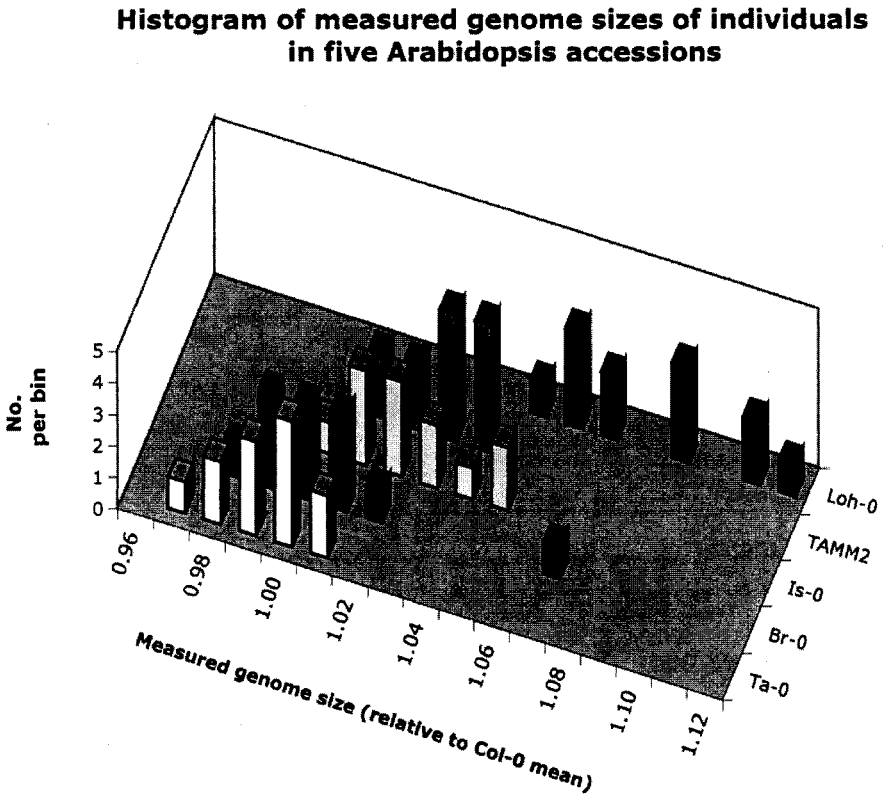


Figure 2.2. Distribution of genome size measurements of five accessions in *Arabidopsis thaliana*. Values given are relative to the average measured genome size of the sequenced accession Col-0. Genome size of three individuals in each accession was assayed by flow cytometry four times over a period of one week.

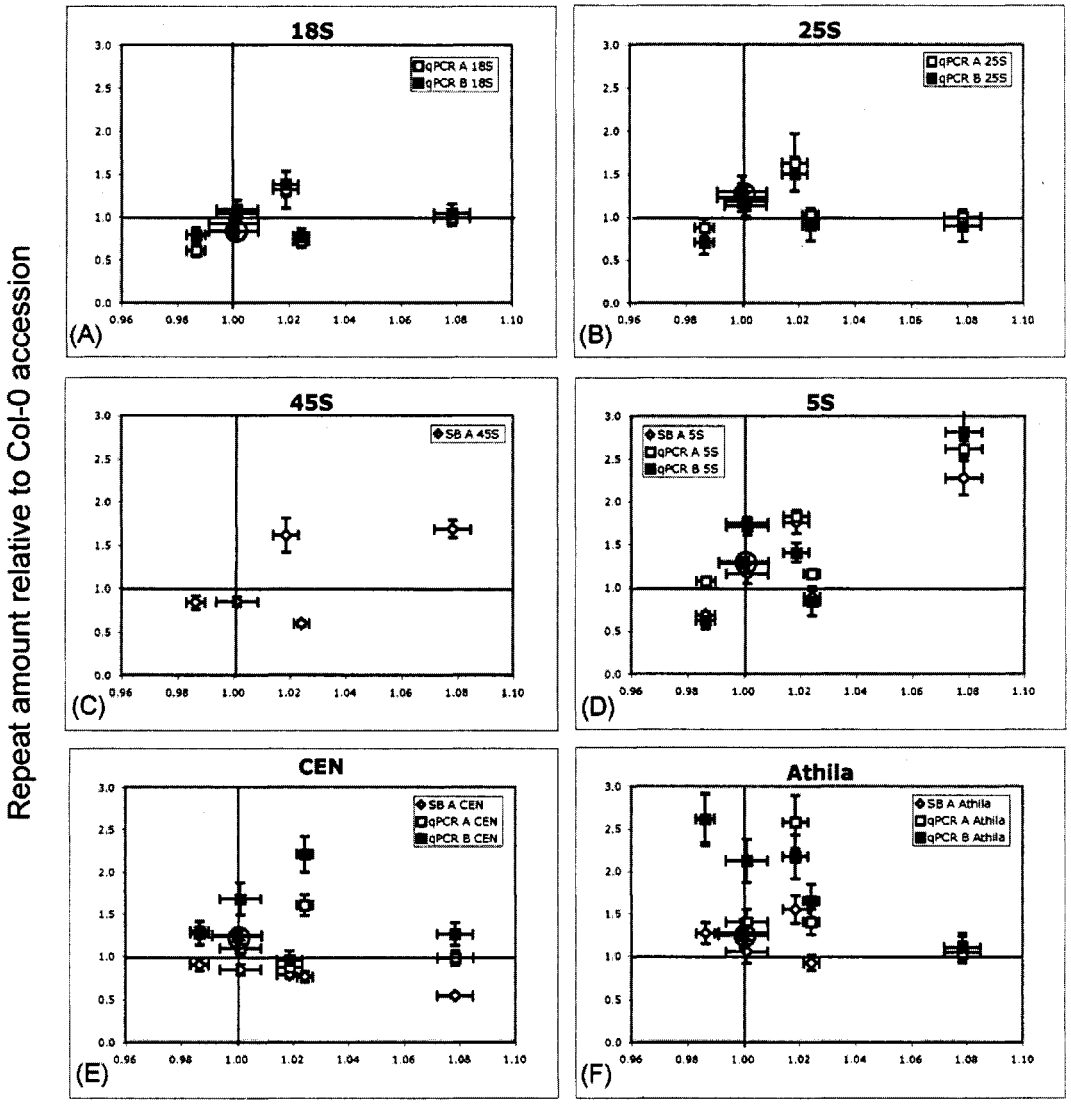
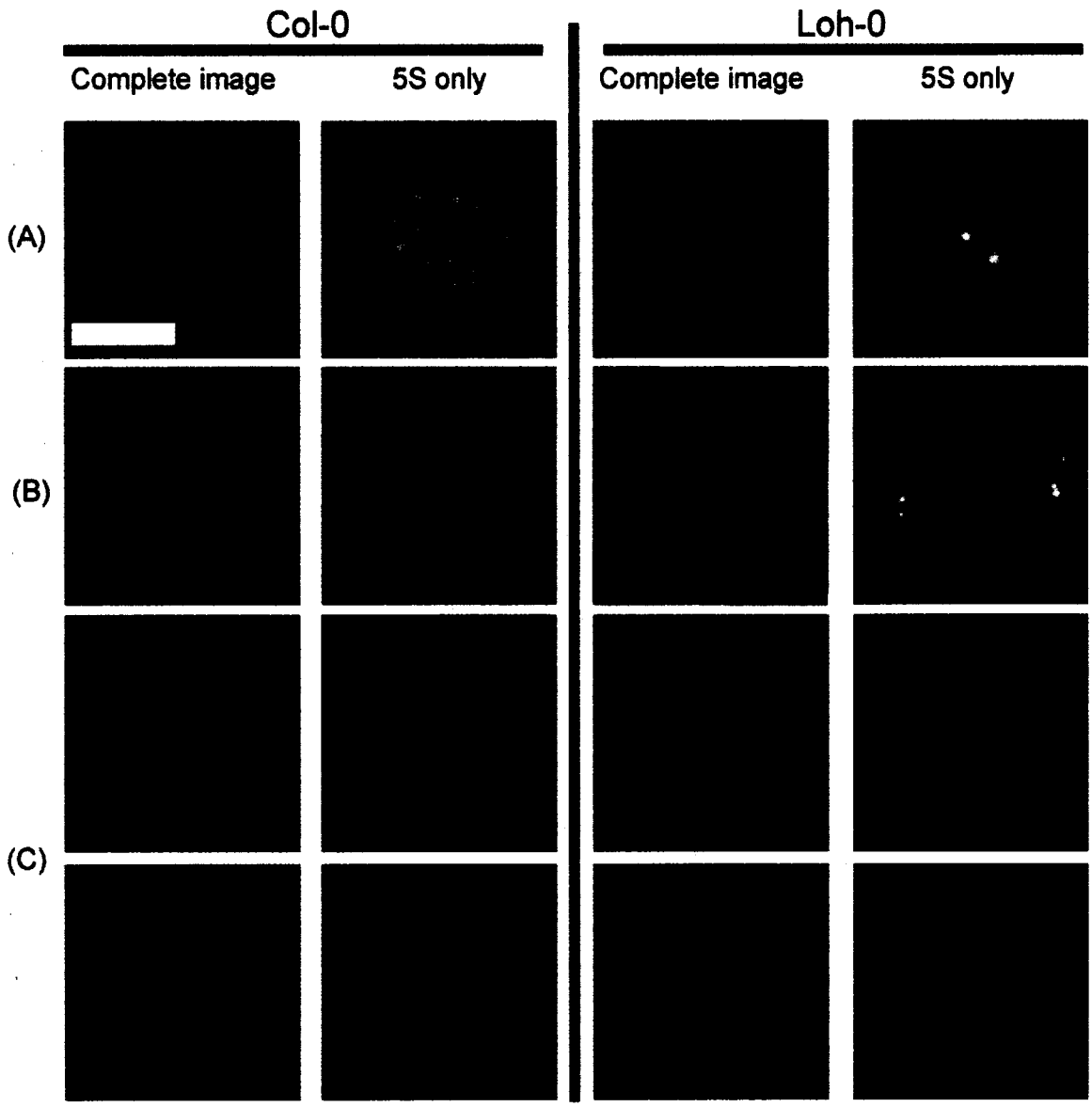
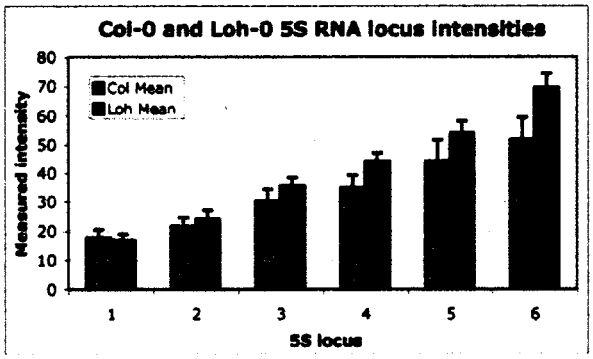


Figure 2.3. Measured sizes of the major heterochromatic repeat arrays in five *Arabidopsis thaliana* accessions, plotted against the accessions' genome sizes as measured by flow cytometry; all values are relative to Col-0. Square markers present the qPCR-measured repeat size; diamonds present slot blot values; white and black markers correspond to the two siblings assayed. The qPCR values from two Col-0 individuals are circled. Horizontal and vertical bars present the standard error of the mean. (A) and (B) qPCR-measured sizes of the 18S and 25S rDNA repeats. The 45S RNA transcript is cleaved into these two subunits. (C) The 45S rDNA array size as measured with slot blot genomic hybridization. (D) 5S rDNA loci. (E) Centromeric 180 bp repeat. (F) Pericentromeric retrotransposon Athila.

Figure 2.4. Fluorescence intensity comparison between the 5S rDNA arrays in the sequenced accession Col-0 and Loh-0, the accession with the largest measured genome size. Images are unmanipulated FISH photomicrographs from anther squashes. For each accession two images are presented: the left is a composite RGB image with green channel showing DAPI stained DNA, blue channel showing the centromere probe, and red channel showing the probe for the 5S rDNA; the right image is the red channel with the 5S hybridization signal alone. The size marker is 10 μm. (A) Col-0 and Loh-0 cells in prophase of meiosis I. (B) Col-0 and Loh-0 cells in anaphase of meiosis I. (C) Two Col-0 and two Loh-0 interphase diploid nuclei. (D) Average intensities of the six 5S loci in Col-0 (11 cells) and Loh-0 (16 cells); spots are first ranked by intensity in each image. Error bars give the standard error of the mean. Protease pretreatment was absent for (A) and (B), and present for (C). Details of the analysis are in Methods and materials.


Col-0
Loh-0
Complete image
5S only
Complete image
5S only
(A)
(B)
(C)

(D)


Col-0 and Loh-0 5S RNA locus intensities
Measured intensity
Col Mean
Loh Mean
0
10
20
30
40
50
60
70
80
1
2
3
4
5
6
5S locus

(A)

Loh-0 included	Specified Col-0 genome (Mbp)	CEN (Mbp)	Athila (Mbp)	45S (Mbp)	5S (Mbp)	Merit function	r	r^2	Regression p-value	RMS difference (Mbp)
Yes	130	8	1	7	6	0.05	0.68	0.46	0.03	3.5
Yes	145	13	1	9	14	0.06	0.78	0.61	0.01	4.6
Yes	160	17	3	19	13	0.09	0.66	0.44	0.04	5.7
No	130	8	1	12	1	0.04	0.70	0.49	0.05	2.5
No	145	15	1	19	2	0.06	0.70	0.49	0.05	4.0
No	160	23	2	19	8	0.09	0.60	0.36	0.11	5.5

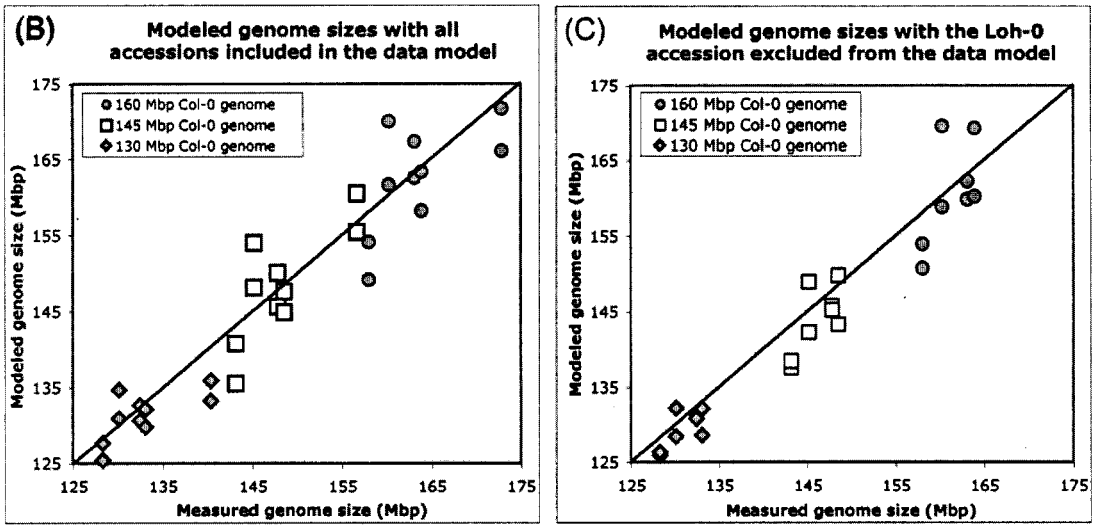


Figure 2.6. The assayed accessions' genome sizes modeled as the sum of a 108 Mbp sequenced basal genome plus four major repeats which vary in size among the accessions. As the absolute size of these repeats is not known even for the Col-0 reference genome we fit them to three possible totals for the *A. thaliana* genome. (A) The modeled size of the repeats in the sequenced genome of Col-0 is given for each of these, along with measures of agreement between the modeled genome sizes and those determined with flow cytometry. (B) Modeled genome size plotted versus the measured genome size for each individual. (C) The same result when the accession with the largest genome is omitted from the model calculations. In (B) and (C) the line of perfect fit between the modeled and measured genomes is drawn.

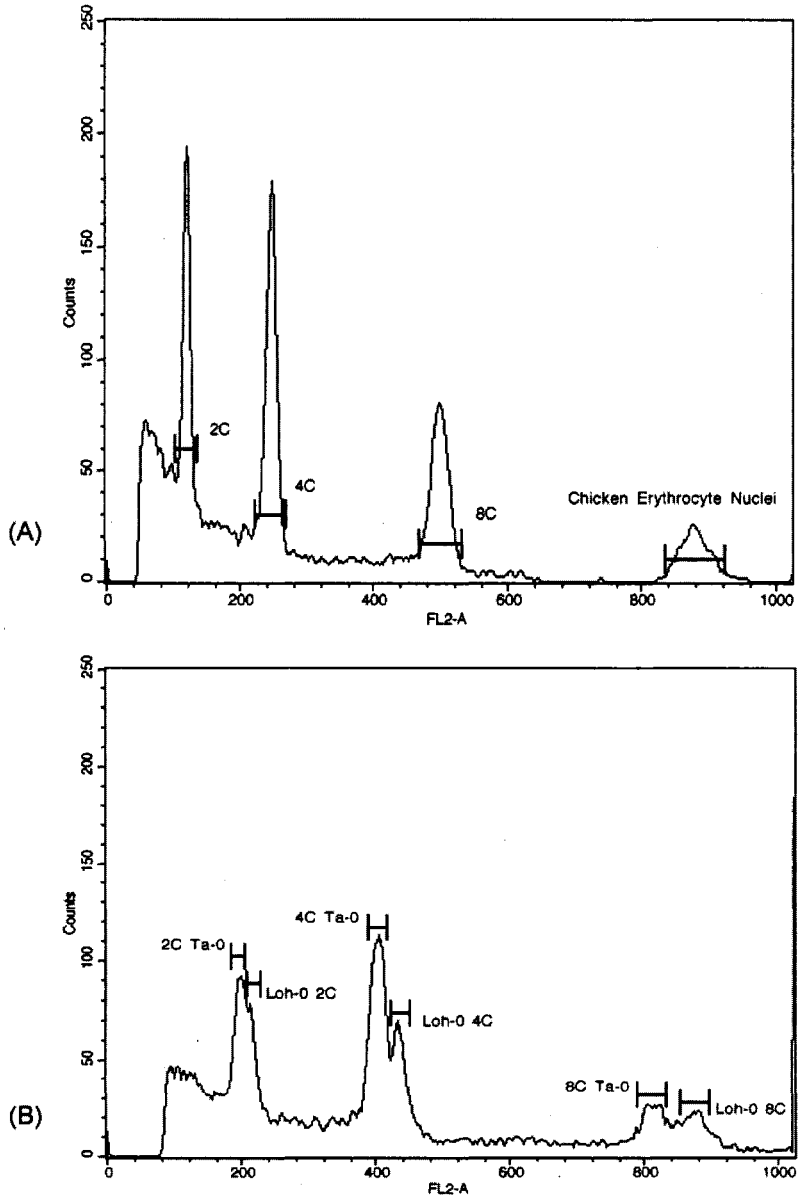
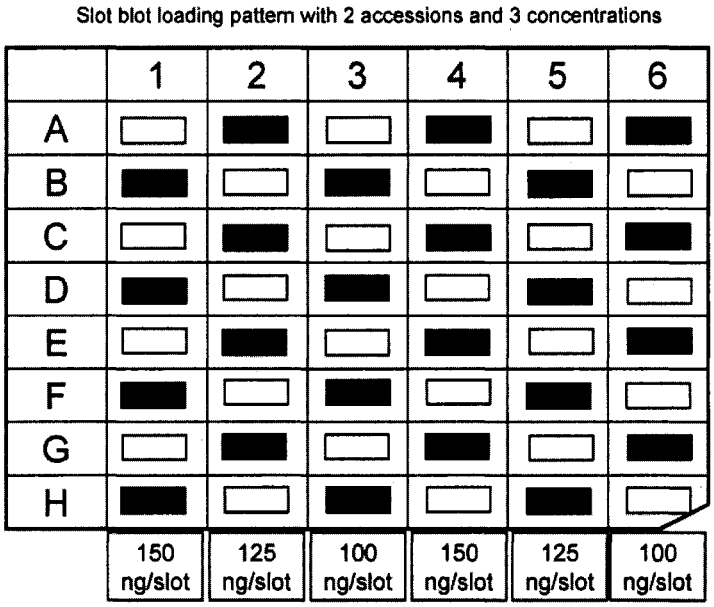


Figure 2.7. Flow cytometry histograms. The fluorescence of propidium iodide-stained nuclei is proportional to DNA content. Intensities are distributed linearly across 1024 channels; a threshold is set to ignore the many particles of low intensity cellular debris. Horizontal scaling differs in the two panels. (A) The relative positions of Arabidopsis diploid (2C) nuclei, endoreplicated 4C and 8C nuclei, and chicken erythrocyte nuclei with known DNA content, permit genome size determination of the Arabidopsis accessions. (B) Leaf samples of two accessions whose measured genome size differs by 20 Mbp (Ta-0 and Loh-0) were prepared together and assayed to produce these sets of double peaks. Ta-0 nuclei are present at a higher concentration than the Loh-0.

(A)



(B)

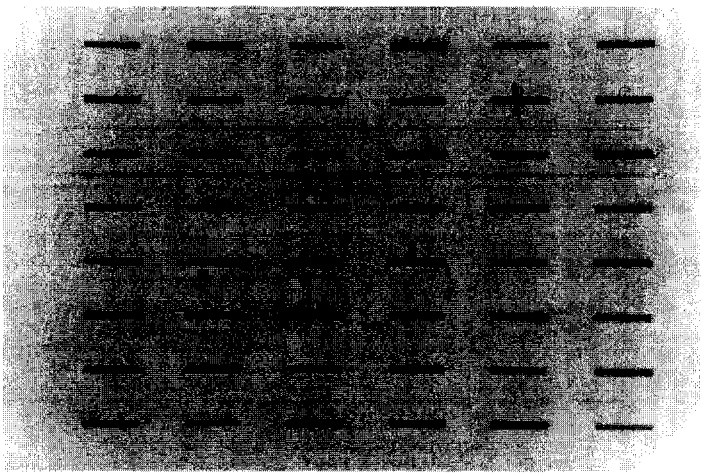


Figure 2.8. Slot blot membrane loading pattern and image of hybridized membrane. (A) Design for loading genomic DNA into a 48-well slot blot apparatus. A single accession's DNA was loaded into 24 slots in three amounts (unshaded slots), and the comparison standard in the remaining (shaded) slots. (B) Representative membrane after probe hybridization and development of the signal.

Chapter III: Replication of heterochromatic repeats

Introduction

Our finding of a negative relationship between measured genome size and the copy number of centromeric repeats (Figure 2.3E) was unexpected, as centromeres may make up the single largest repeat class in the genome (Figure 2.6A). While the result is based on just five accessions and is strongly affected by the single accession with the largest measured genome, the slot blot data are highly significant. The centromere-associated Athila transposon shows the same negative correlation of copy number and genome size (Figure 2.3F). We were led by these counter-intuitive findings to consider whether differential replication between euchromatin and heterochromatin in Arabidopsis could be biasing our measurements; I report in this chapter the work we did to assess that question.

As clearly illustrated in Figure 2.7A, nuclei in the leaves of Arabidopsis are present in several discrete ploidy levels. Such polyploidization occurs in both plants and animals (Larkins *et al.*, 2001) and is known as endoreduplication or endoreplication. Flow cytometry enables the determination of the distribution of the endoreplicated nuclei in a plant sample (see Figure 2.7A). The distribution of ploidy levels is not constant within the plant body; in particular leaves and unopened flower buds differ markedly, as shown in Table 3.1.

Heterochromatic repeats have been demonstrated to be substantially under-replicated in the endoreplicated cells of some taxa, for example in Drosophila nurse cells (Hammond and Laird, 1985). We investigated whether under-replication of repeated heterochromatic sequences in the higher ploidy nuclei of Arabidopsis distorts our genome size measurements or our assessment of the amounts of heterochromatin in the accessions. Our findings, however, appear to be relevant to whole-genome replication.

Differential endoreplication

Differential replication of chromatin in cycles of endoreplication was observed in Drosophila (Rudkin, 1969) and later quantified for five classes of chromatin in the same species by Hammond and Laird (1985). In large nurse cells (mitotic sister cells of oocytes and replicated up to 2048C) the authors found satellite DNA (a centromere marker) present at 4.5% of that expected under full replication, histone genes at 35%, telomere sequences at 63%, 28S and 18S rDNA at 90%, and 5S rDNA at 102%. Nurse cells differ from salivary gland cells, also endoreplicated, in that banded polytene chromosomes are not seen in large nurse cells. Both under- and over-replication of chromatin have been observed in other taxa (Hammond and Laird, 1985).

A plant sample generally contains cells of several ploidy levels, and if differential replication occurs the measured fraction of heterochromatin will not be the same as in diploid nuclei. I generated a computer model demonstrating this effect; Figure 3.1 presents the outcome for a leaf with the ploidy distribution in Table 3.1. Taking an example, the leftmost solid line, labeled 0.1, is the measured value of 10% heterochromatin. At the top of the chart, at 100% heterochromatin replication, the line is coincident with the vertical line placed at a 0.1 fraction of 2C heterochromatin; that is, there is no distortion of heterochromatin content. As less of the heterochromatin is replicated, descending the chart, the solid line of 10% measured heterochromatin curves to the right. At around 15% heterochromatin replication, its 2C fraction is 20%, while what will be measured in leaf material using slot blots or qPCR is 10%.

It is also possible that under-replication becomes more pronounced as the amount of heterochromatin increases. Indeed, models of under-replication in Drosophila propose that heterochromatin suppresses its own replication in endoreplicated cells: replication forks stall within the condensed chromatin as the S-phase concludes (Laird, 1973). Additional amounts of contiguous or intercalary heterochromatin, expected with increased levels of centromeric or

pericentromeric repeats, arguably increase the level of under-replication. This would be consistent with our own measurements of decreasing centromeric repeat copy number with increasing flow cytometry-determined genome size.

Results

Flow cytometry with 2C, 4C, 8C nuclei

Under-replication of heterochromatin in leaf nuclei is potentially observable using flow cytometry. The locations of histogram peaks will be shifted to the left relative to their position under complete genome doubling. These displacements provide a measure of the unreplicated fraction; the 4C peak will be displaced by 4 times the difference from full doubling and the 8C peak by 8 times that value. Divergence (d) of the 8C/4C intensity ratio from exactly 2.0 is related to the amount differentially replicated (y) and the size of the full genome (x) by the equation $d = 2y/x$. The distance between peaks rather than peak positions is used, as cytometers often do not accurately measure absolute peak locations; there is typically a non-zero intercept in the linear relationship between nuclear DNA base pair amount and its fluorescence. For that reason we can only assess under-replication beginning with the 4C ploidy level.

Table 3.2 presents the ratios of the separation of the 8C and 4C peaks divided by the separation of the 4C and 2C peaks, in the five accessions assayed in detail with flow cytometry in chapter two. Four of the six ratios are significantly different from the value of 2.0 expected under complete doubling, as is the average of all values, and indicate some amount of under-replication.

Comparative Genomic Hybridization of leaf vs bud

The differences in ploidy distribution between leaf and floral bud cells present an opportunity to measure under-replication in Arabidopsis. Because buds are typically 75% 2C, and leaves are 66% 4C and higher, leaf DNA will show the effect of under-replication more strongly than bud DNA, if it occurs, and providing

that under-replication is happening similarly in both organs. In the case of under-replication, the ratios of heterochromatin to single copy genes will vary systematically in the DNA obtained from the two organs.

Figure 3.2 presents results of comparative genomic hybridization (CGH) assays similar to those presented in Figure 2.5. In this case we are comparing DNA extracted from leaf to DNA extracted from flower bud, in the same individual. Panel (A) assays the reference Col-0 accession, panel (B) accession Loh-0. The values plotted are feature hybridization intensities of leaf DNA divided by bud intensities, or L/B ratios. We found these ratios for the approximately 20,000 features ranged from less than 0.5 to over 2.0, a factor of four. For comparison, the range of ratios in the Col-0 bud vs bud control assay of Figure 2.5C was 0.8 to 1.2. No heterochromatic repeats are represented on the arrays; nonetheless a slight reduction of the intensity of leaf hybridization relative to bud is present pericentromerically in both accessions. These results reveal differences in gene copy number in the two organs of the same plant.

To assess whether the variation in the two datasets of Figure 3.2 is due to a process shared between the two accessions, we generated a scatterplot of the two sets of ratios (Figure 3.3A). The two are positively correlated, with 42% of the variance in their L/B ratios shared between them, indicating changes in copy number along the chromosomes tend to occur at the same positions in the two accessions. Because the arrays do not directly assess heterochromatin, the result addresses only variation in copy number of the genes in Arabidopsis, between leaf and bud. The scatterplot of Col-0 L/B ratios against the B/B Col-0 control assay, Figure 3.3B, shows that greater variation is present in the former and that what remains in the self-self assay is uncorrelated with it ($r^2 = 0.04$). The covariation thus is likely of biological origin and not an artifact of the microarray method.

Quantitative PCR of leaf vs bud

We used quantitative PCR (qPCR) to test the CGH finding of Figure 3.3A, and also to measure any differential replication of the heterochromatic repeats in the two organs. First, we selected a single feature on each chromosome where the L/B ratio of both accessions in Figure 3.2 was either substantially less than or more than one. We developed primers for fragments of exons of these 10 genes (Table 3.3) and amplified them in leaf- and bud-extracted DNA in six accessions. The single copy genes ACT2 and ROC1 were also amplified in order to measure relative genome copy number in the DNA extracts from leaf and bud. The resulting L/B ratios are given in Table 3.4. The same results for qPCR of the heterochromatic repeats are presented in Table 3.5. Overall the qPCR L/B gene copy ratios are not significantly different from one, for both the "low" and "high" genes of Table 3.3, averaged across all accessions. The heterochromatin L/B ratios of Table 3.5 are more variable, but the averages across all accessions also are not significantly different from one.

Discussion

Table 3.2 provides evidence for under-replication of sequences in endoreplicated leaf nuclei; four of the six accessions' 8C/4C intensity ratios are significantly different from exact doubling (p-value < 0.05), as is the average of all measurements. The deviation from doubling, however, is less than one percent overall and amounts to about 1.4 Mbp in a genome of 156 Mbp. From the slot blot hybridization data, the quantity of centromeric repeats in accession Col-0 is about 7 Mbp more than that in Loh-0. The under-replication observed is then insufficient by a factor of five to account for the negative relationship between measured genome size and centromeric repeat amount.

While the comparative hybridization results may indirectly indicate some under-replication of pericentromeric heterochromatin, the significant covariation in L/B ratios between genic sequences (Figure 3.3) in the two surveyed

accessions appears to relate to a genome-wide replication process. Prokaryotes typically replicate the genome with a single pair of replication forks moving in opposite directions from the origin of replication, but eukaryote replication requires a large collection of these replication units, called replicons, arranged in tandem along the chromosomes. In Arabidopsis replicons are approximately 50,000 base pairs long (Van't Hof *et al.*, 1978); because average gene separation is one tenth of that, the CGH assays have the resolution to sample a replicon multiple times.

As flower bud cells are rapidly dividing, Figure 3.3 reflects genome replication begun and interrupted by collection of the material. CGH in that case has the potential to identify the origins of replication in Arabidopsis. Resolution may be weak, especially if the order in which origins of replication fire is random.

The qPCR results presented in Table 3.4 do not support the CGH findings. Across all six accessions, the "low" feature mean L/B ratio is lower than the "high" feature three out of five times, but the mean copy number ratio of leaf to bud across all accessions is close to one for each feature. For these 10 genes there is little or no difference in copy number in leaf and bud. It may be the analytical power of the gene array lies in the ability to sample 20,000 features simultaneously, with lower confidence in a single value; sampling less than 0.05 percent of the features to verify that result with a more quantitative method is too few. Yet because the qPCR results are clear this explanation is not entirely satisfactory and the contradiction remains.

Looking at the qPCR-determined ratios in any one accession across all features, the three accessions with the smallest measured genome size have L/B ratios greater than one, the next two (including Col-0) have ratios less than one, and the Loh-0 mean is greater than one. Some of these distributions significantly differ from one another, for example Lip-0 and Col-0, and all accessions' means are significantly different from one. These systematic differences could potentially be explained by errors in estimates of genome copy number in the separate leaf and bud DNA extracts of an accession; if either were in error the ratios could be

biased. However, the two independent qPCR measurements of copy number in samples using different single-copy genes agreed well (r^2 = 0.98, linear regression p-value ~ 10^{-8}) across all samples. Thus, genome copy number estimates are probably accurate.

The analogous results for the heterochromatic repeats in Table 3.5 have an overall mean not significantly different from one, but as with the 10 genes discussed above the average L/B value in a single accession is significantly different from one, except for Lip-0 and Nok-3. Col-0 and Loh-0 have similar heterochromatin L/B means and they exhibit L/B values for 5S rDNA of 0.30 or less, which is substantially lower than the corresponding values for the other accessions (0.81 to 1.30). The correlation between the repeat L/B values of these two accessions is 0.98. We noted in chapter two that heterochromatin copy numbers of Col-0 and Loh-0 are quite similar, except for the 5S rDNA; the two accessions may be closely related.

One finding from the qPCR assays is particularly interesting and suggestive. The correlation between the 25S and 18S rDNA copy number in the 12 samples is 0.79 (regression p-value < 0.01). As mentioned previously, this is not surprising since they are both subcomponents of the 45S rDNA. The correlation in copy number between the 5S and 25S rDNA is small, 0.13 (regression p-value = 0.68); there is no *a priori* reason to expect that the two would change in copy number together as accessions diverge. But while agreement of the L/B ratios for 25S and 18S across the accessions is high, with r = 0.89 (regression p-value = 0.02), the correlation between the 5S and 25S L/B ratios is an equally impressive 0.92 (regression p-value = 0.01). Potentially then the same process or management activity is modulating copy number change from bud to leaf in these non-contiguous rDNA fractions.

While the qPCR results suggest differential (both over and under) replication of heterochromatin in Arabidopsis, they are confounded by equally variable gene L/B ratios. The latter finding and the covariation between two accessions in the CGH data suggest the occurrence of wholesale replication in

progress when the plant material was collected. The CGH data separately suggest the possibility of using this method to identify the origins of replication in Arabidopsis.

Table 3.1. Distribution of endoreplicated nuclei in different Arabidopsis organs. The fraction of total nuclei in endoreplication states as determined with flow cytometry.

Ploidy of nucleus	Leaf fraction in this ploidy	Flower bud fraction in this ploidy
2C	0.33	0.75
4C	0.35	0.20
8C	0.22	0.05
16C	0.10	0.00

Table 3.2. Flow cytometry-determined fraction of full doubling between 4C and 8C endoreplicated nuclei. Because of non-zero intercepts (see text) this is calculated as (8C - 4C)/(4C - 2C) where 2C, 4C and 8C are the fluorescences measured for nuclei of those endoreplication states. Relative MGS is the genome size of the accession relative to the reference accession Col-0. Means in bold type are significantly different from 2.0 at the 5% confidence level as they are more than two standard errors from 2.0.

Accn	Rel. MGS	Fractional change in fluorescence from 4C to 8C peak	Standard error of the mean	Deficit in genome doubling (Mbp)	N
Ta-0	0.99	**1.985**	0.006	1.2	12
Br-0	1.00	2.004	0.005	-0.3	12
Col-0	1.00	**1.969**	0.007	2.4	23
Is-0	1.02	**1.973**	0.008	2.1	26
TAMM-2	1.02	**1.982**	0.003	1.4	12
Loh-0	1.08	1.997	0.006	0.3	12
	All values averaged	**1.982**	0.003	1.4	97

Table 3.3. CGH assay features selected for qPCR measurements. The ten features on the Arabidopsis oligo array selected to reassess with a second method, the CGH findings of differences in gene copy number in DNA extracted from leaves (L) and buds (B) from the same individual. One feature with L/B ratio greater than 1.0 and one feature with L/B ratio less than 1.0 were selected on each chromosome.

Chrom. No.	Position	Gene ID	qPCR ID	Col L/B ratio	Loh L/B ratio
1	3095302	At1g09570	X1 lo	0.69	0.60
1	2752186	At1g08650	X1 hi	1.52	1.71
2	14496761	At2g34520	X2 lo	0.51	0.46
2	15134870	At2g36210	X2 hi	1.53	2.30
3	1883680	At3g06220	X3 lo	0.55	0.47
3	3991270	At3g12580	X3 hi	1.53	2.17
4	3330360	At4g07530	X4 lo	0.72	0.72
4	14349947	At4g31805	X4 hi	1.48	2.07
5	18917059	At5g47320	X5 lo	0.69	0.67
5	22151731	At5g55330	X5 hi	1.60	1.56

Table 3.4. qPCR leaf to bud (L/B) gene copy number ratios in six Arabidopsis accessions. qPCR was used to measure the copy number of 10 genes in DNA separately extracted from leaves and flower buds in single individuals. Values given are the ratio of the copy number in the leaf-extracted DNA to that extracted from buds (L/B). The genes are identified in Table 3.2. In the two rightmost columns the mean L/B ratio and standard error of the mean for the 10 genes in each accession is given; the same parameters for each gene in the 6 accessions is given in the bottom two rows. Means in bold type are significantly different from 1.0 at the 5% confidence level as they are more than two standard errors from 1.0.

Accn	X1 lo	X1 hi	X2 lo	X2 hi	X3 lo	X3 hi	X4 lo	X4 hi	X5 lo	X5 hi		
Ta-0	1.04	1.06	1.00	1.06	1.00	1.05	1.22	1.13	1.09	1.15	**1.08**	0.02
Lip-0	1.01	1.20	1.12	1.59	1.17	1.19	1.19	0.97	0.99	1.13	**1.16**	0.06
Br-0	0.98	1.08	1.06	1.15	1.02	1.10	1.15	1.16	1.06	1.11	**1.09**	0.02
Col-0	0.90	0.93	0.86	1.04	0.99	0.87	1.03	0.83	0.81	0.88	**0.91**	0.03
Nok-3	1.04	0.96	0.91	0.99	1.02	0.93	0.96	0.95	0.89	0.87	**0.95**	0.02
Loh-0	1.12	1.06	1.06	1.19	1.00	0.97	1.16	1.17	0.89	0.97	**1.06**	0.03
	1.02	1.05	1.00	1.17	1.03	1.02	**1.12**	1.03	0.95	1.02	**Mean**	
	0.03	0.04	0.04	0.09	0.03	0.05	0.04	0.06	0.05	0.05		**Stderr**

Table 3.5. Leaf to bud copy number ratios for major genomic repeats. The ratio of copy number in leaves and buds (L/B) of several heterochromatic repeats in six Arabidopsis accessions as determined by qPCR. Means in bold type are significantly different from 1.0 at the 5% confidence level as they are more than two standard errors from 1.0.

Accn	CEN	Athila	5S	25S	18S		
Ta-0	0.95	1.09	1.30	1.48	1.18	**1.20**	0.09
Lip-0	1.11	0.84	0.81	0.87	0.94	0.91	0.05
Br-0	1.12	1.11	1.21	1.19	1.03	**1.13**	0.03
Col-0	0.81	0.97	0.24	0.74	0.67	**0.69**	0.12
Nok-3	0.84	0.95	0.94	1.11	0.85	0.94	0.05
Loh-0	0.86	1.03	0.30	0.78	0.61	**0.71**	0.12
	0.95	1.00	0.80	1.03	0.88	**Mean**	
	0.06	0.04	0.18	0.12	0.09		**Stderr**

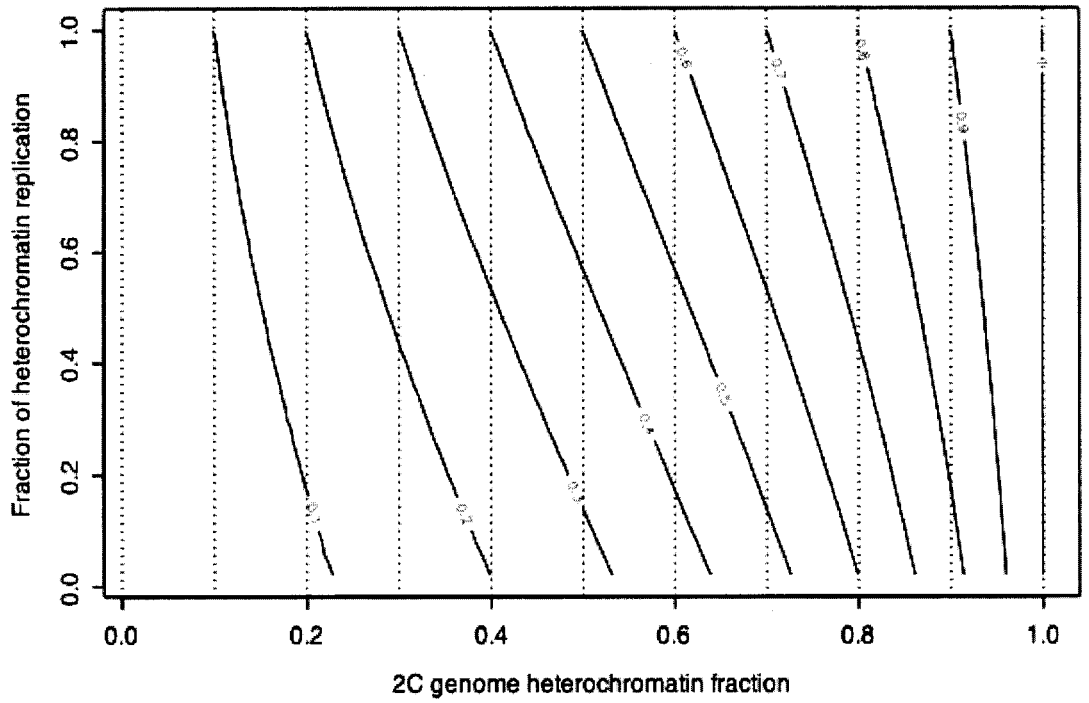


Figure 3.1. Computational model demonstrating the effect of any differential endoreplication in Arabidopsis on its measurement in leaves. The horizontal axis represents the heterochromatin fractions of possible 2C genomes; values range from zero to one, that is, no heterochromatin or 100% heterochromatin. The vertical axis is the fraction of heterochromatin that is replicated at each endoreplication of a chromosome. The possibility ranges from zero to one; the latter value represents complete doubling. The solid lines in the chart present the fraction of heterochromatin that is measured; dotted vertical lines serve as a guide.

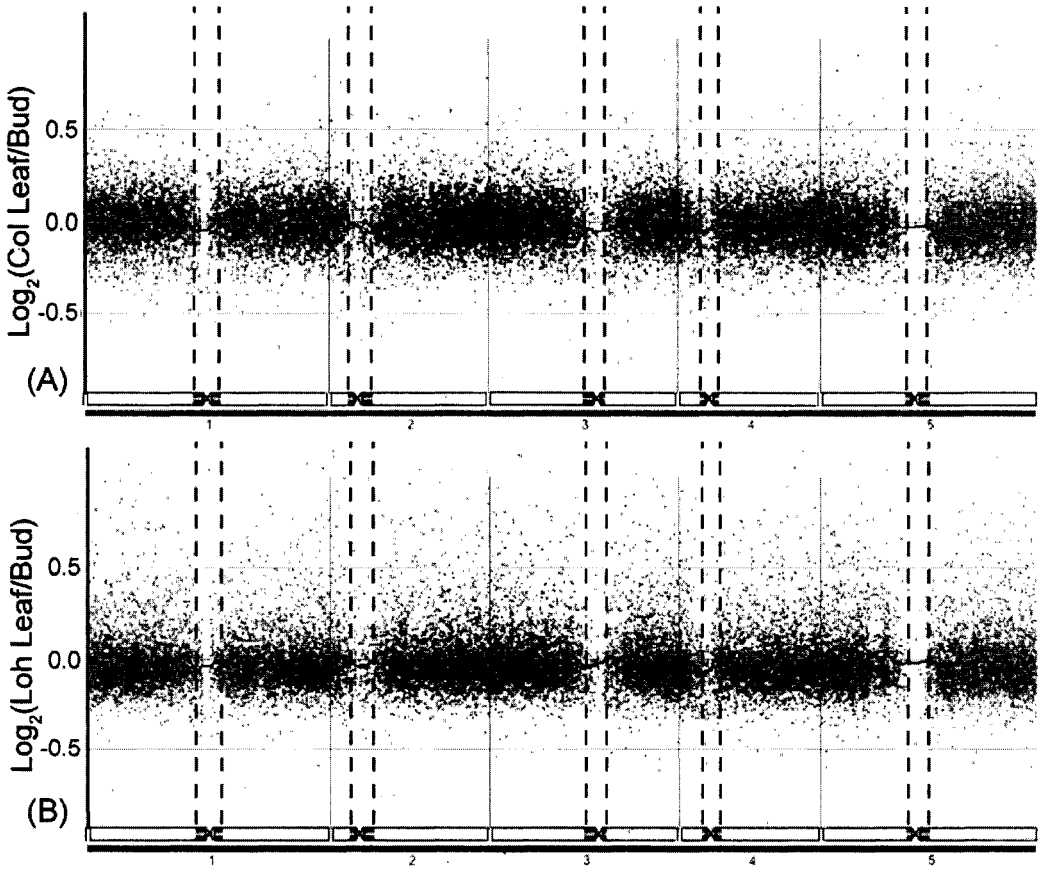
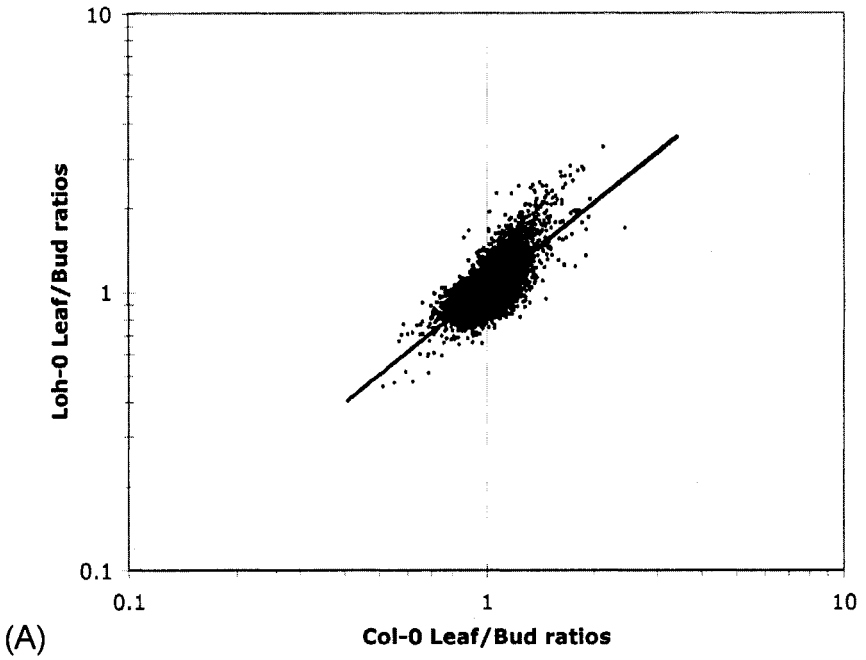
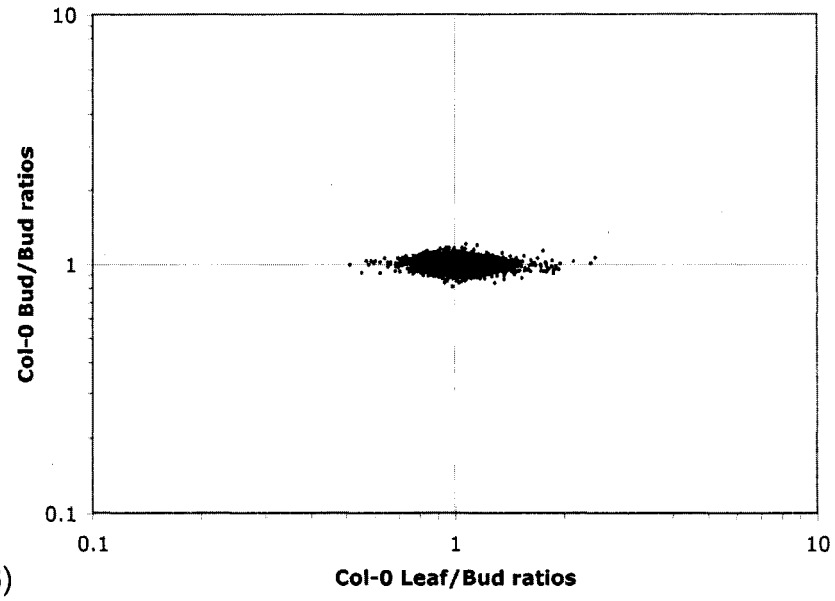


Figure 3.2. Two CGH arrays hybridized with leaf and bud DNA. Results are formatted as in Figure 2.5B. Values are the hybridization intensity of leaf-extracted genomic DNA divided by the intensity of flower bud-extracted DNA, from the same individual. The solid line is a 101-point running average. (A) With DNA extracted from the Col-0 accession (B) DNA extracted from Loh-0, the accession with the largest measured genome size in this study.



(A)



(B)

Figure 3.3. Scatterplots of CGH array ratios. (A) The leaf/bud hybridization ratios of Figure 3.2A, Col-0 accession, plotted against the same ratios of Figure 3.2B, Loh-0 accession. Variation in leaf/bud ratios along the chromosomes in the two accessions is positively correlated, with $r^2 = 0.42$. The linear trendline between the two datasets is given. (B) A similar plot comparing variation in Col-0 leaf/bud ratios with Col-0 bud/bud values (the self-self comparison of Figure 2.5C). The two have a negative correlation near zero, with $r^2 = 0.04$.

Chapter IV: Conclusions

Determining the complete genome of multicellular organisms remains a goal rather than a common accomplishment. The genomic variation among individuals, and within a single organism over time and in its parts, are in general unknown. Speaking broadly, the more detailed genomic information becomes available on members of a population, the more variation is found; and while some types and degrees of genomic variation have no observable effect (Rogers and Bendich, 1987; Beaulieu, 1992), others are known or thought to be extremely significant (Redon *et al.*, 2006). Until accurate, rapid and cheap methods are developed to identify complete genomes and their polymorphisms in many individuals, the genomic era's goal of separating significant differences from the unimportant, and understanding the determining mechanisms, may not be met.

The approach used in this study, measuring total genome size in individuals, quantifying their variation in a set of known repeats, and assessing with a computational model how well the latter values can account for the former, is partly successful. Several limitations are clear: the set of assayed repeats was small and any variation in unknown repeats could not be identified. The methods used have coarse resolution, qPCR for example able to identify differences approaching 20% at best. The reasons for agreement and disagreement between this method and slot blot hybridization (the correlation coefficient between them for the 5S rDNA arrays is 0.96; for the 180-bp centromeric repeat, 0.29) were not identified. The model itself should be tested with fictional datasets to better understand its capabilities and weaknesses. Nonetheless the approach is valid in that the total genome must equal the sum of its parts, when both pieces can be sufficiently well identified.

Other methods show promise to reduce the approach's limitations. Genomic *in situ* hybridization (GISH) could be used to demonstrate the presence of repeats that are specific or enriched for a given genome. For example, one

might label total genomic DNA from the Loh-0 accession with a fluorescent tag and use it to stain the chromosomes of Loh-0 and of Col-0; unlabeled competitor Col-0 DNA is added in excess to the hybridization mix to blocks hybridization of common sequences. Preferential staining by the labeled probe of Loh-0 chromosomes would reveal repeated elements that are differentially enriched or unique in the Loh-0 genome, although it would not provide a direct identification. If present, such elements would be likely to contribute to the extra size of the Loh-0 genome.

The gold standard for determination of genomic variation nonetheless remains complete processive sequencing of the genome, not possible today, but efforts proceed. Massively parallel sequencing methods can now be applied to a genome yielding low cost sequencing through the high throughput sequencing of small genomic fragments (Margulies *et al.*, 2005; Shendure *et al.*, 2005). These methods can be used to provide a relatively unbiased measurement of sequence copy number in a genomic fragment set (Crawford *et al.*, 2006; Pinard *et al.*, 2006). The acquisition of such signatures from fragmented genomic DNA of the Col-0 and Loh-0 accessions would have the potential to quantify the copy number of both known and unknown repeats.

Signature methods of sequencing may be adequate to infer copy number of repeated elements, but the short length of each signature represents a limitation and assembly of repeated regions would be exceedingly difficult. An alternative sequencing method, direct linear analysis (DLA), under development, reads fluorescent tags placed on single DNA molecules within a nanofluidic mechanism and it may eventually be adapted to sequence long repeated arrays (Chan *et al.*, 2004; Phillips *et al.*, 2005). DNA is introduced into the device by pipetting and forced through narrowing photolithographically-defined channels; lasers scan fluorescent sequence-specific DNA tags as the stretched molecule passes through a detector. Its average read rate is estimated at 1 Mbp per second. The system has shown applicability to mapping by identifying distances

between specific tags, but requires a difficult 1000-fold increase in resolution to sequence each base pair.

All considered, the problem of elucidating the full genome sequence and composition is a challenging one. The methods I employed have considerable power and worked satisfactorily. As new methods become possible and affordable, a more complete and more satisfactory picture of the Arabidopsis genome should emerge.

Bibliography

AGI. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796-815.

Aitman, T. J., Dong, R., Vyse, T. J., Norsworthy, P. J., Johnson, M. D., Smith, J., et al. (2006). Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*, *439*(7078), 851-855.

Atkin, N. B. & Brito-Babapulle, V. (1981). Heterochromatin polymorphism and human cancer. *Cancer Genet Cytogenet*, *3*(3), 261-272.

Beaulieu, G. C. (1992). *Copy number variability in four multigene families of maize.* University of Washington, Seattle.

Bennett, M. D. (1998). Plant genome values: how much do we know? *Proc Natl Acad Sci U S A*, *95*(5), 2011-2016.

Bennett, M. D., Leitch, I. J., Price, H. J., & Johnston, J. S. (2003). Comparisons with Caenorhabditis (approximately 100 Mb) and Drosophila (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb. *Ann Bot (Lond)*, *91*(5), 547-557.

Bennetzen, J. L. (2000). The many hues of plant heterochromatin. *Genome Biol*, *1*(1), reviews1.07.1-107.4.

Bennetzen, J. L. (2002). Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica*, *115*(1), 29-36.

Bino, R. J., Lanteri, S., Verhoeven, H. A., & Kraak, H. L. (1993). Flow Cytometric Determination of Nuclear Replication Stage in Seed Tissues. *Ann Bot*, *72*(2), 181-187.

Bubner, B. & Baldwin, I. T. (2004). Use of real-time PCR for determining copy number and zygosity in transgenic plants. *Plant Cell Rep*, *23*(5), 263-271.

Campell, B. R., Song, Y., Posch, T. E., Cullis, C. A., & Town, C. D. (1992). Sequence and organization of 5S ribosomal RNA-encoding genes of

Arabidopsis thaliana. *Gene*, *112*(2), 225-228.

Capparelli, R., Cottone, C., D'Apice, L., Viscardi, M., Colantonio, L., Lucretti, S., et al. (1997). DNA content differences in laboratory mouse strains determined by flow cytometry. *Cytometry*, *29*(3), 261-266.

CESC. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science*, *282*(5396), 2012-2018.

Chan, E. Y., Goncalves, N. M., Haeusler, R. A., Hatch, A. J., Larson, J. W., Maletta, A. M., et al. (2004). DNA mapping using microfluidic stretching and single-molecule detection of fluorescent site-specific tags. *Genome Res*, *14*(6), 1137-1146.

Clarke, B., McKay, I., Grigliatti, T., Lloyd, V., & Yuan, A. (1996). A Markov model for the assembly of heterochromatic regions in position effect variegation. *J Theor Biol*, *181*(2), 137-155.

Cloix, C., Tutois, S., Mathieu, O., Cuvillier, C., Espagnol, M. C., Picard, G., et al. (2000). Analysis of 5S rDNA arrays in Arabidopsis thaliana: physical mapping and chromosome-specific polymorphisms. *Genome Res*, *10*(5), 679-690.

Cloix, C., Tutois, S., Yukawa, Y., Mathieu, O., Cuvillier, C., Espagnol, M. C., et al. (2002). Analysis of the 5S RNA pool in Arabidopsis thaliana: RNAs are heterogeneous and only two of the genomic 5S loci produce mature 5S RNA. *Genome Res*, *12(1)*(1), 132-144.

Comai, L., Tyagi, A. P., & Lysak, M. A. (2003). FISH analysis of meiosis in Arabidopsis allopolyploids. *Chromosome Res*, *11*(3), 217-226.

Comeron, J. M. (2001). What controls the length of noncoding DNA? *Curr Opin Genet Dev*, *11*(6), 652-659.

Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., et al. (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res*, *16*(1), 123-131.

Cremer, T. & Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet*, *2*(4), 292-301.

Creusot, F., Fouilloux, E., Dron, M., Lafleuriel, J., Picard, G., Billault, A., et al. (1995). The CIC library: a large insert YAC library for genome mapping in Arabidopsis thaliana. *Plant J*, *8*(5), 763-770.

CSHL. (2000). The complete sequence of a heterochromatic island from a higher eukaryote. The Cold Spring Harbor Laboratory, Washington University Genome Sequencing Center, and PE Biosystems Arabidopsis Sequencing Consortium. *Cell*, *100*(3), 377-386.

Darzynkiewicz, Z., Traganos, F., Kapuscinski, J., Staiano-Coico, L., & Melamed, M. R. (1984). Accessibility of DNA in situ to various fluorochromes: relationship to chromatin changes during erythroid differentiation of Friend leukemia cells. *Cytometry*, *5*(4), 355-363.

Dernburg, A. F., Sedat, J. W., & Hawley, R. S. (1996). Direct evidence of a role for heterochromatin in meiotic chromosome segregation. *Cell*, *86*(1), 135-146.

Dillon, N. & Festenstein, R. (2002). Unravelling heterochromatin: competition between positive and negative factors regulates accessibility. *Trends Genet*, *18*(5), 252-258.

Dimitri, P. & Junakovic, N. (1999). Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin. *Trends Genet*, *15*(4), 123-124.

Dolezel, J., Greilhuber, J., Lucretti, S., Meister, A., Lysak, M. A., Nardi, L., et al. (1998). Plant Genome Size Estimation by Flow Cytometry: Inter-laboratory Comparison. *Annals of Botany*, *82*(Supplement 1), 17-26.

Eichler, E. E., Clark, R. A., & She, X. (2004). An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat Rev Genet*, *5*(5), 345-354.

Feschotte, C., Jiang, N., & Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics. *Nat Rev Genet*, *3*(5), 329-341.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, *269*(5223), 496-512.

Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., et al. (2004). Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol*, *2*(7), E207.

Fransz, P., Armstrong, S., Alonso-Blanco, C., Fischer, T. C., Torres-Ruiz, R. A., & Jones, G. (1998). Cytogenetics for the model system Arabidopsis thaliana. *Plant J*, *13*(6), 867-876.

Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M., et al. (2006). Copy number variation: New insights in genome diversity. *Genome Res*, *16*(8), 949-961.

Gentle, A., Anastasopoulos, F., & McBrien, N. A. (2001). High-resolution semi-quantitative real-time PCR without the use of a standard curve. *Biotechniques*, *31*(3), 502, 504-506, 508.

Giangare, M. C., Prosperi, E., Pedrali-Noy, G., & Bottiroli, G. (1989). Flow cytometric evaluation of DNA stainability with propidium iodide after histone H1 extraction. *Cytometry*, *10*(6), 726-730.

Gregory, T. R. & Hebert, P. D. (1999). The modulation of DNA content: proximate causes and ultimate consequences. *Genome Res*, *9*(4), 317-324.

Hammond, M. P. & Laird, C. D. (1985). Chromosome structure and DNA replication in nurse and follicle cells of Drosophila melanogaster. *Chromosoma*, *91*(3-4), 267-278.

Haupt, W., Fischer, T. C., Winderl, S., Fransz, P., & Torres-Ruiz, R. A. (2001). The centromere1 (CEN1) region of Arabidopsis thaliana: architecture and functional impact of chromatin. *Plant J*, *27*(4), 285-296.

Henry, I. M., Dilkes, B. P., & Comai, L. (2006). Molecular karyotyping and aneuploidy detection in A. thaliana using quantitative fluorescent PCR. *Plant J*, *48*(2), 307-319.

Hillier, L. W., Miller, W., Birney, E., Warren, W., Hardison, R. C., Ponting, C. P., et al. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, *432*(7018), 695-716.

Hosouchi, T., Kumekawa, N., Tsuruoka, H., & Kotani, H. (2002). Physical map-

based sizes of the centromeric regions of Arabidopsis thaliana chromosomes 1, 2, and 3. *DNA Res*, *9*(4), 117-121.

Iborra, F. J. & Cook, P. R. (2002). The interdependence of nuclear structure and function. *Curr Opin Cell Biol*, *14*(6), 780-785.

Ishii, K., Arib, G., Lin, C., Van, H. G., & Laemmli, U. K. (2002). Chromatin boundaries in budding yeast: the nuclear pore connection. *Cell*, *109*(5), 551-562.

Jackson, D. A. (1991). Structure-function relationships in eukaryotic nuclei. *Bioessays*, *13*(1), 1-10.

Jackson, M. S., Rocchi, M., Thompson, G., Hearn, T., Crosier, M., Guy, J., et al. (1999). Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications and unstable sequences with homologies to telomeric and other centromeric locations. *Hum Mol Genet*, *8*(2), 205-215.

Ji, Y., Eichler, E. E., Schwartz, S., & Nicholls, R. D. (2000). Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res*, *10*(5), 597-610.

Josefsson, C., Dilkes, B., & Comai, L. (2006). Parent-dependent loss of gene silencing during interspecies hybridization. *Curr Biol*, *16*(13), 1322-1328.

Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*, *16*(9), 418-420.

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, *115*(1), 49-63.

Klug, W., & Cummings, M. (2004). *Essentials of Genetics* (5th ed.). Prentice Hall.

Kubis, S., Schmidt, T., & js, H.-H. (1998). Repetitive DNA Elements as a Major Component of Plant Genomes. *Ann Bot (Lond)*, *82*, 45-55.

Kumekawa, N., Hosouchi, T., Tsuruoka, H., & Kotani, H. (2000). The size and sequence organization of the centromeric region of arabidopsis thaliana chromosome 5. *DNA Res*, *7*(6), 315-321.

Kumekawa, N., Hosouchi, T., Tsuruoka, H., & Kotani, H. (2001). The size and sequence organization of the centromeric region of Arabidopsis thaliana chromosome 4. *DNA Res*, *8*(6), 285-290.

Laird, C. D. (1973). DNA of Drosophila chromosomes. *Annu Rev Genet*, *7*, 177-204.

Lamond, A. I. & Earnshaw, W. C. (1998). Structure and function in the nucleus. *Science*, *280*(5363), 547-553.

Larionov, A., Krause, A., & Miller, W. (2005). A standard curve based method for relative real time PCR data processing. *BMC Bioinformatics*, *6*, 62.

Larkins, B. A., Dilkes, B. P., Dante, R. A., Coelho, C. M., Woo, Y. M., & Liu, Y. (2001). Investigating the hows and whys of DNA endoreduplication. *J Exp Bot*, *52*(355), 183-192.

Levin, D. A. (2002). *The Role of Chromosomal Change in Plant Evolution (Oxford Series in Ecology and Evolution)* (ISBN: 0195138600 ed.). Oxford Press.

Lewin, B. (1997). *Genes VI.* Oxford University Press, USA.

Lingjaerde, O. C., Baumbusch, L. O., Liestol, K., Glad, I. K., & Borresen-Dale, A. L. (2005). CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics*, *21*(6), 821-822.

Lohe, A. R. & Hilliker, A. J. (1995). Return of the H-word (heterochromatin). *Curr Opin Genet Dev*, *5*(6), 746-755.

Manuelidis, L. (1990). A view of interphase chromosomes. *Science*, *250*(4987), 1533-1540.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376-380.

McClintock, B. (1930). A cytological demonstration of the location of an interchange between two non-homologous chromosomes of *Zea Mays*. *PNAS*, *16*, 791-796.

McClintock, B. (1933). The association of non-homologous parts of

chromosomes in the mid-prophase of meiosis in *Zea mays*. *Z. Zellforsch. Mikrosk. Anat*, *19*, 191-237.

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science*, *226*, 792-801.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., et al. (2000). A whole-genome assembly of Drosophila. *Science*, *287*(5461), 2196-2204.

Noirot, M., Barre, P., Louarn, J., Duperray, C., & Hamon, S. (2000). Nucleus-Cytosol Interactions--A Source of Stoichiometric Error in Flow Cytometric Estimation o f Nuclear DNA Content in Plants. *Annals of Botany*, *86*(2), 309-316.

Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., et al. (2005). The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol*, *3*(7), e196.

Park, J. P., Wojiski, S. A., Spellman, R. A., Rhodes, C. H., & Mohandas, T. K. (1998). Human chromosome 9 pericentric homologies: implications for chromosome 9 heteromorphisms. *Cytogenet Cell Genet*, *82*(3-4), 192-194.

Peirson, S. N., Butler, J. N., & Foster, R. G. (2003). Experimental validation of novel and conventional approaches to quantitative real-time PCR data analysis. *Nucleic Acids Res*, *31*(14), e73.

Petrov, D. A. (2001). Evolution of genome size: new approaches to an old problem. *Trends Genet*, *17*(1), 23-28.

Phillips, K. M., Larson, J. W., Yantz, G. R., D'Antoni, C. M., Gallo, M. V., Gillis, K. A., et al. (2005). Application of single molecule technology to rapidly map long DNA and study the conformation of stretched DNA. *Nucleic Acids Res*, *33*(18), 5829-5837.

Pinard, R., de Winter, A., Sarkis, G. J., Gerstein, M. B., Tartaro, K. R., Plant, R. N., et al. (2006). Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*, *7*, 216.

Press, W., Flannery, B., Teukolsky, S., & Vetterling, W. (1986). Modeling of Data.

Numerical Recipes, the Art of Scientific Computing. Cambridge: Cambridge University Press.

Puertas, M. J. (2002). Nature and evolution of B chromosomes in plants: A non-coding but information-rich part of plant genomes. *Cytogenet Genome Res*, *96*(1-4), 198-205.

Redi, C. A., Garagna, S., Zacharias, H., Zuccotti, M., & Capanna, E. (2001). The other chromatin. *Chromosoma*, *110*(3), 136-147.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), 444-454.

Rogers, S. O. & Bendich, A. J. (1987). Heritability and variability in ribosomal RNA genes of *Vicia Faba. Genetics*, *117*, 285-295.

Rudkin, G. T. (1969). Non replicating DNA in Drosophila. *Genetics*, *61*(1), Suppl:227-38.

Rutledge, R. G. & Cote, C. (2003). Mathematics of quantitative kinetic PCR and the application of standard curves. *Nucleic Acids Res*, *31*(16), e93.

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., et al. (2003). TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, *34*(2), 374-378.

Schmuths, H., Meister, A., Horres, R., & Bachmann, K. (2004). Genome Size Variation among Accessions of Arabidopsis thaliana. *Ann Bot (Lond)*.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., et al. (2004). Large-scale copy number polymorphism in the human genome. *Science*, *305*(5683), 525-528.

Shabtai, F. & Halbrecht, I. (1979). Risk of malignancy and chromosomal polymorphism: a possible mechanism of association. *Clin Genet*, *15*(1), 73-77.

Sharma, T., Bardhan, A., & Bahadur, M. (2003). Reduced meiotic fitness in hybrids with heterozygosity for heterochromatin in the speciating Mus terricolor complex. *J Biosci*, *28*(2), 189-198.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., et al. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, *309*(5741), 1728-1732.

Suda, J. (2004). *An employment of flow cytometry into plant biosystematics.* Charles University, Prague.

Szymanski, M., Barciszewska, M. Z., Erdmann, V. A., & Barciszewski, J. (2003). 5 S rRNA: structure and interactions. *Biochem J*, *371*(Pt 3), 641-651.

TAIR. (2005). Sequence Viewer. Retrieved 5 July, 2006, from http://www.arabidopsis.org/servlets/sv

TAIR. (2006). Genome Assembly. Retrieved 5 July, 2006, from http://www.arabidopsis.org/portals/genAnnotation/gene_structural_annotation/agicomplete.jsp

Temsch, E. M. & Greilhuber, J. (2001). Genome size in Arachis duranensis: a critical study. *Genome*, *44*(5), 826-830.

TIGR. (2000). BAC Tiling Path. Retrieved 5 July, 2006, from http://www.tigr.org/tigr-scripts/euk_manatee/listchromosomes.cgi?db=ath1

Van't Hof, J., Kuniyuki, A., & Bjerknes, C. A. (1978). The size and number of replicon families of chromosomal DNA of Arabidopsis thaliana. *Chromosoma*, *68*(3), 269-285.

Vieira, C., Nardon, C., Arpin, C., Lepetit, D., & Biemont, C. (2002). Evolution of genome size in Drosophila. is the invader's genome being invaded by transposable elements? *Mol Biol Evol*, *19*(7), 1154-1161.

Weimarck, A. (1975). Heterochromatin polymorphism in the rye karyotype as detected by the giemsa C-banding technique. *Hereditas*, *79*(2), 293-300.

West, A. G., Gaszner, M., & Felsenfeld, G. (2002). Insulators: many functions, many mechanisms. *Genes Dev*, *16*(3), 271-288.

Wortman, J. R., Haas, B. J., Hannick, L. I., Smith, R. K. J., Maiti, R., Ronning, C. M., et al. (2003). Annotation of the Arabidopsis genome. *Plant Physiol*, *132*(2), 461-468.

Yakin, K., Balaban, B., & Urman, B. (2005). Is there a possible correlation between chromosomal variants and spermatogenesis? *Int J Urol*, *12*(11), 984-989.

Vita

Jerry Davison earned a B.S. and M.S. in Physics at the University of Missouri. At the University of Washington he earned a B.S. in Botany and a Ph.D. in Biology.